DOCUMENT RESUME

ED 402 571 CS 012 689

| | |
|---|---|
| TITLE | Quality and Utility: The 1994 Trial State Assessment in Reading. The Fourth Report of the National Academy of Education Panel on the Evaluation of the NAEP Trial State Assessment: 1994 Trial State Assessment in Reading. |
| INSTITUTION | National Academy of Education, Stanford, Calif. |
| SPONS AGENCY | Department of Education, Washington, DC. |
| REPORT NO | ISBN-0-942469-09-7 |
| PUB DATE | 96 |
| NOTE | 221p. |
| PUB TYPE | Reports - Research/Technical (143) |
| EDRS PRICE | MF01/PC09 Plus Postage. |
| DESCRIPTORS | *Data Analysis; Elementary Secondary Education; Limited English Speaking; Program Design; *Reading Achievement; *Reading Research; *Test Use; *Test Validity |
| IDENTIFIERS | *National Assessment of Educational Progress; *Trial State Assessment (NAEP) |

ABSTRACT

This report evaluates the conduct, validity, and uses of the National Assessment of Educational Progress (NAEP) Trial State Assessment (TSA). The report addresses such pressing problems as how participation in NAEP can be maintained and appropriate samples can be achieved; how errors can be minimized in the complex process of scaling and analyzing data; how the definition of achievement levels can be accomplished; how inclusion of children with limited English proficiency or disabilities can be included and reported; how private schools can be included and reported; and how the NAEP state assessments relate to the national NAEP. After an introduction, sections of the report are The Content Validity of the 1994 Reading Assessment; Sampling and Assessment Administration for the 1994 TSA; The assessment of Students with Disabilities or Limited English Speaking Proficiency; Scaling and Analysis of the 1994 Reading Assessment; Reading Achievement Levels; Reporting and Dissemination for the 1994 Reading Assessment; and Conclusions and Recommendations. Contains 66 references, and 21 tables and 7 figures of data. Appendixes present detailed scoring guides and examples of student responses for sample assessment times shown in figure 2.1; reading experts participating in the panel's content validity study for the 1994 TSA; and synopses of studies for the National Academy of Education Panel on the Evaluation of the National Assessment of Educational Progress Trial State Assessment. (RS)

# *Quality and Utility:*
## *The 1994 Trial State Assessment in Reading*

*The Fourth Report of*
*The National Academy of Education Panel*
*on the Evaluation of the NAEP Trial State Assessment:*
*1994 Trial State Assessment in Reading*



*Panel Chairmen*
Robert Glaser, University of Pittsburgh
Robert Linn, University of Colorado

*Project Director*
George Bohrnstedt, American Institutes for Research

# *Quality and Utility:*
## *The 1994 Trial State Assessment in Reading*

The contents of this book were developed
under a grant from the Department of Education.
However, those contents do not
necessarily represent the policy of the
Department of Education and you
should not assume endorsement by the Federal Government.


ISBN 0-942469-09-7

Printed on recycled paper in the United States of America
The National Academy of Education
Stanford University
School of Education CERAS-108
Stanford, CA 94305-3084
(415) 725-1003
Designed and printed by Armadillo Press, Mountain View, CA

# *Quality and Utility:*
## *The 1994 Trial State Assessment in Reading*

*The Fourth Report of*
*The National Academy of Education Panel*
*on the Evaluation of the NAEP Trial State Assessment:*
*1994 Trial State Assessment in Reading*

*Panel Chairmen*
Robert Glaser, University of Pittsburgh
Robert Linn, University of Colorado

*Project Director*
George Bohrnstedt, American Institutes for Research

# The National Academy of Education

The National Academy of Education is composed of scholars and education leaders who "promote scholarly inquiry and discussion concerning the ends and means of education, in all its forms, in the United States and abroad." Our current active membership is limited to 125 scholars. The heart of the Academy is found in the lively discussions that take place in our regular meetings and in the special panels and committees that we establish. Throughout our 31-year history the Academy has been called upon by governmental and other agencies to conduct special studies and reviews on education issues of public interest, ranging from desegregation to the teaching of reading to standards-based reform. During the past six years, a panel of the Academy has been monitoring, studying, and making recommendations concerning the conduct of the Trial State Assessments given since 1990 in conjunction with the National Assessment of Educational Progress. For the first time in the history of NAEP, these assessments allow state-by-state comparisons of education achievement; they have proven to be of great interest to educators. The Academy's Panel has prepared three in-depth biennial reports on the trial state assessments of 1990, 1992, and, in this report, 1994. They also issued, at the request of the National Center for Education Statistics, which oversees the Panel's work, a special report on the setting of achievement levels in connection with NAEP.

To carry out the Panel's challenging assignment, the Academy engaged two outstanding education researchers, Robert Glaser and Robert Linn, as co-chairs, and George Bohrnstedt and his colleagues at the American Institutes of Research as subcontractors to assist in the conduct of the research and writing, as well as a panel of distinguished educators and researchers with diverse forms of expertise on the NAEP program. They have constructed and overseen an ongoing set of research and policy papers examining major issues concerning the state trials of NAEP, including validity, sampling, content, data analysis, and reporting issues. In the coming year, the Panel will conclude its work with a capstone report that will reflect on broad, long-term issues regarding the future of state NAEP and its relationship to national NAEP.

Carl F. Kaestle
President, The National Academy of Education

# Table of Contents

# Detailed Table of Contents

April 29, 1996

Jeanne Griffith
Acting Commissioner
National Center for Education Statistics
U.S. Department of Education
555 New Jersey Avenue 20208-5653

Dear Jeanne:

On behalf of The National Academy of Education, I am pleased to transmit to you the fourth report of the Academy's Panel on the Evaluation of the NAEP Trial State Assessment, entitled *Quality and Utility: The 1994 Trial State Assessment in Reading*. In it, the Panel evaluates the conduct, validity, and uses of that assessment. They apply their guiding principles and their research findings to many policy issues concerning the future of the NAEP state assessments.

This report has been reviewed and approved by The National Academy of Education's Executive Council, acting as a Committee of Readers. Like the preceding reports of this panel, we are confident that it will be helpful to policymakers in reaching thoughtful decisions about important policy issues concerning the National Assessment of Educational Progress.

This report addresses such pressing problems as how participation in NAEP can be maintained and appropriate samples be achieved; how errors can be minimized in the complex process of scaling and analyzing the data; how the definition of achievement levels can be accomplished; how inclusion of children with limited English proficiency or disabilities can be included and reported; how private schools can be included and reported; and how the NAEP state assessments relate to the national NAEP.

In addition to these and other consequential issues surrounding the ongoing administration and interpretation of state NAEP, there are a number of longer-term issues associated with the future of state and national NAEP and the assessment's relationship to developments in American education. To address these issues, the Panel plans shortly to issue a capstone report. Issues surrounding the redesign of NAEP stem from fundamental changes occurring in the field of educational assessment and from the ongoing education reform movement, aimed at transforming the content and structure of American education. To determine NAEP's most constructive role in the midst of these changes is both very difficult and very important. The Academy is proud of the contributions the Panel has made to that effort over the past six years, and we look forward to its capstone report as the culminating advice of this distinguished group of educators and scholars.

Sincerely,

Carl F. Kaestle

Carl F. Kaestle
President, The National Academy of Education
Professor of Education, The University of Chicago

# *Acknowledgments*

This mandated report presents the findings of The National Academy of Education concerning its independent evaluation of the 1994 NAEP Trial State Assessment (TSA).

The Panel once again acknowledges the indispensable support provided by staff at the National Center for Education Statistics. Information, practical assistance, and reasoned advice have all been offered freely throughout the course of the evaluation, greatly facilitating the work of the Panel. In particular, retired Commissioner Emerson Elliott and his immediate successor, Acting Commissioner Jeanne Griffith, both provided thoughtful guidance at many project junctures, while Garry Phillips, Sharif Shakrani, and Larry Ogle all have played very active roles in securing required technical materials and keeping the Panel apprised of NAEP plans and activities. Ed Mooney has continued to assist the staff with the administrative details of the project.

The courteous and professional collaboration of the NAEP contractors has also enriched the Panel's evaluation and smoothed its progress. Among the many Educational Testing Service (ETS) staff who contributed to the Panel's work on the 1994 TSA were Eugene Johnson, who provided extensive information and advice concerning NAEP's psychometric procedures; Jay Campbell, who answered many questions about the reading assessment and also coordinated the special scoring of student responses from the Panel's studies of accessibility and exclusions among students with disabilities or limited English proficiency; Dave Freund and Patricia O'Reilly, who oversaw the preparation of data sets for the Panel's analyses; and Debbie Kline, who coordinated responses to the Panel's questions about the 1994 Technical Report. John Mazzeo of ETS also played a key role by answering questions on a wide range of topics and coordinating response to the Panel's diverse requests.

Staff at Westat and National Computer Systems (NCS) have also given freely of their advice and assistance concerning NAEP's sampling, administration, and data processing procedures. Special thanks are due to Nancy Caldwell and Diane Walsh of Westat for facilitating the recruitment of schools for the Panel's studies of assessibility and exclusions, observations of TSA assessment sessions for the Panel's study of the assessment administration, and our attendance at Westat training sessions, as well as for generally helping to coordinate the Panel's field activities with the NAEP schedule. Keith Rust of Westat served as an essential resource concerning NAEP's sampling design and implementation, and Brad Thayer of NCS expedited the flow of administration documents needed for the Panel's studies of assessibility and exclusions.

Thanks are also due to the National Assessment Governing Board (NAGB) staff, including Roy Truby, Mary Lyn Bourque, and Ray Fields, for their cooperation and prompt response to requests for information. Sue Loomis and other American College Testing (ACT) staff graciously welcomed Panel observers at the 1994 standard setting sessions, the revisit of the reading achievement levels, and meetings of the Technical Advisory Committee for Standard Setting, as well as promptly fulfilling our requests for information. Members of the Assessment Subcommittee, Education Information Advisory Committee, and Council of Chief State School Officers also welcomed our representatives, while Cadell Hemphill coordinated agendas and kept Panel staff informed of Subcommittee activities. Mary Baronne of ETS played a similar liaison role with respect to meetings and activities of the NAEP Design and Analysis Committee.

The Panel expresses its appreciation to the state assessment directors and curriculum specialists who responded to its various surveys on NAEP administration and on the utility and impact of NAEP reports, to state department of education employees who provided us with state assessment data for the Panel's study of public school nonparticipation, and to the many students and teachers who cooperated with our studies of assessibility and exclusions.

The Panel extends thanks to the staff at American Institutes for Research for their role in drafting and redrafting the various chapters that comprise this report. Special thanks goes to Fran Stancavage who took the lead in drafting most of the chapters. Others contributing heavily to the writing and editing were George Bohrnstedt, Jennifer O'Day, Evelyn Hawkins, John Olson, and Lorna Bennie.

Finally, the Panel extends its gratitude to the distinguished researchers who carried out the commissioned investigations. Thanks are due to principle investigators P. David Pearson (Michigan State University), Lizanne DeStefano (University of Illinois, Urbana-Champaign), Bruce Spencer (Northwestern University), Larry Hedges (University of Chicago), and Richard Venezky (University of Delaware), as well as to James Yesseldyke, Kevin McGrew, and Martha Thurlow (National Center on Educational Outcomes) who provided consultation for the Panel's study of assessibility and exclusions. Summaries of the commissioned reports are included in appendix C, this volume, while the full text of the reports are contained in volume two.

# Foreword

Since 1990, every cycle of the National Assessment of Educational Progress (NAEP) has included an option for states to participate on a voluntary basis and receive state-level results in at least one subject area at one grade level. State NAEP assessments were first authorized by Congress in 1988, at which time Congress mandated that an evaluation of the feasibility and technical adequacy of such assessments be conducted for trials in 1990 and 1992. Pursuant with this legislation, Trial State Assessments (TSAs) were conducted in eighth-grade mathematics in 1990 and in fourth- and eighth-grade mathematics and fourth-grade reading in 1992, and Congress subsequently extended the trials and the evaluation to include the 1994 assessment as well. This report has been prepared in response to that mandate and provides an evaluation of the 1994 Trial State Assessment (TSA) in fourth-grade reading by The National Academy of Education's Panel on the Evaluation of the Trial State Assessment.

The Panel's work on the 1994 TSA fourth-grade reading assessment has taken place in a period during which numerous innovations in assessment have been implemented at the national, state, and local levels. In this context, NAEP serves as a valuable independent monitor of status and trends for student achievement as our nation proceeds toward improved education for all children and youth. However, NAEP too has changed and, in order to be effective, must continue to change and adapt to the many requirements posed by new content, new techniques for measuring performance, and more inclusive coverage of the nation's diversity. The Panel believes that systematic study of such innovations and their results should continue to be an essential part of efforts to enhance our nation's key independent indicator of educational progress.

This is the fourth of the Panel's reports. Encompassing numerous studies and analytical papers commissioned by the Panel, these reports have served to inform technical and policy issues under consideration by Congress, the National Center for Education Statistics (NCES), the National Assessment Governing Board (NAGB), and the NAEP contractors.

The first report, *Assessing Student Achievement in the States*, was issued in March, 1992. It presented the Panel's findings and observations on the first TSA, which was conducted in eighth-grade mathematics in 1990. More specifically, the report presented the Panel's observations on the assessment's content validity, sampling, administration, scoring and interpretation, as well as on the reporting of results to the public and press. While the Panel concluded that the trial was largely a success, a set of recommendations for changes was also included in the report.

The second report, *Setting Performance Standards for Student Achievement,* issued in September, 1993, studied the new set of performance standards, called achievement levels, that were being implemented for reporting and interpreting NAEP results. The Panel's report examined the process used for setting achievement levels, the validity and reasonableness of the 1992 achievement levels in reading and mathematics, and the relationship of NAEP to emerging national education standards. The Panel's report, we believe, has made a valuable contribution to the continuing discussion and debate about how best to set performance standards on an assessment such as NAEP.

*The Trial State Assessment: Prospects and Realities*, the Panel's third report, was issued in December, 1993 and examined the 1992 state assessments in reading and mathematics as well as critical questions surrounding the continuance of a state NAEP. In addition to issues of sampling and administration, content validity, and reporting, this report presented a set of guiding principles that could inform, not only the recommendations made in the report, but also discussions and decisions concerning the TSA made by Congress, NCES, and NAGB.

The Panel's forthcoming capstone report will be released in fall, 1996 and will address the role of NAEP in education reform and choices that confront NAEP now and as we approach the 21st century. Among the latter are choices about how NAEP can best incorporate modern understandings of the acquisition and organization of knowledge, exploit new technologies, accommodate individuals with special needs, and link with other assessment and other educationally relevant data sets to provide richer information on the progress of American education.

Robert Glaser, Chair
Robert Linn, Co-Chair

George Bohrnstedt, Project Director

April 1996

# *The National Academy of Education Panel on the Evaluation of the NAEP Trial State Assessment Project*

**Robert Glaser, Chairman**
Director, Learning Research and Development Center (LRDC)
and National Research Center on Student Learning
University of Pittsburgh

**Robert Linn, Chairman**
Co-director, National Center for Research
on Evaluation, Standards, and Student Testing
University of Colorado at Boulder

**Anthony Alvarado**
Community Superintendent
Community School District 2
NYC Board of Education

**Gordon M. Ambach**
Executive Director,
Council of Chief State School Officers

**Lloyd Bond**
Professor, Educational Research
Methodology
University of North Carolina at
Greensboro

**Ann Brown**
Professor of Education in Math, Science,
and Technology
Evelyn Lois Corey Fellow in Instructional
Science
University of California at Berkeley

**Alonzo Crim**
Professor of Education
Spelman College

**Pasquale J. DeVito**
Rhode Island Department of Education

**Edmund W. Gordon**
The John M. Musser Professor of
Psychology, Emeritus, Yale University
and Distinguished Professor of
Educational Psychology,
City University of New York

**Robert Groves**
Professor of Sociology and Associate
Director
Joint Program in Survey Methodology
University of Michigan

**Richard Jaeger**
Excellence Foundation Professor of
Education and Director,
Center for Educational Research and
Evaluation
University of North Carolina at
Greensboro

**Lyle Jones**
Research Professor
L.L. Thurstone Psychometric Laboratory
Department of Psychology
The University of North Carolina at
Chapel Hill

**Mary M. Lindquist**
Fuller E. Callaway Professor of
Mathematics Education
Columbus State University

**P. David Pearson**
Michigan State University

**Edward Roeber**
Director of Student Assessment Programs
Council of Chief State School Officers

**Albert Shanker**
President
American Federation of Teachers

**Lorrie M. Shepard**
Professor of Education
University of Colorado at Boulder

**Lauress Wise**
President, Human Resources Research
Organization (HumRRO)

*Project Staff*
*American Institutes for Research*

George Bohrnstedt, Project Director
Fran Stancavage, Associate Project Director
Jill Allen
Lorna Bennie
Michelle M. Bullwinkle
Phyllis DuBois
Dey Ehrlich
Catherine Godlewski
Elizabeth Hartka
Evelyn Hawkins
David Huang
Elise McCandless
Don McLaughlin
Jennifer O'Day
John Olson
Marianne Perie
Inna Shapotina
Audrey Struve
Jin-Ying Yu

*Learning Research and Development Center*

Elizabeth Rangel
Cindy Yockel

# Executive Summary

## *Quality and Utility: The 1994 Trial State Assessment in Reading*

The National Assessment for Educational Progress (NAEP) has been the nation's leading indicator of academic achievement for more than 25 years, providing fair and accurate information about the performance of U.S. students in core subject areas. With findings based on representative samples of students in grades 4, 8, and 12, NAEP has long been recognized as an unparalleled resource for educators, policy makers, and all others concerned with national trends in educational progress.

Analyses of NAEP trend data in the past decades revealed a significant narrowing of the achievement gap between African American and white students during the late 1970s and 1980s, while subsequent changes in these same data patterns alerted educators and policy makers to an apparent reopening of that gap in the 1990s. Similarly, NAEP has pointed to important trends in specific subject areas, documenting, for example, the declining achievement of U.S. students in science between 1970 and 1990, as well as the limited time spent on science instruction in most elementary classrooms. These latter data have provided evidential support for groups such as the American Association for the Advancement of Science and the National Science Foundation, who have argued for greater attention to science education in American schools. Finally, NAEP data were also cited in discussions and debates that, in 1989, led the governors and then, in 1994, Congress to target improved science achievement as an important national education goal.

In 1988, NAEP's role was substantially expanded. Responding to increased education reform activity and heightened interest in monitoring progress within the states, Congress lifted the prohibition against collecting and reporting NAEP data at the state level. Public Law 100-297 authorized voluntary state NAEP assessments on a trial basis for 1990 and 1992; this authorization was subsequently extended to include a third trial state assessment (TSA) in 1994. In recognition of the significance of the state NAEP experiment, Congress also called for an independent evaluation of the TSA to judge its feasibility, quality, and utility. Under a grant from the National Center for Education Statistics (NCES), The National Academy of Education (NAE) established this Panel to undertake the evaluation.

Three previous Panel reports, spanning the first two TSAs, have been submitted to Congress. In them, the Panel concluded that the TSAs were successful and should be continued. Areas for further study were also identified, including areas in which the full consequences would not be evident before the trials were scheduled to end. Accordingly the Panel, in its most recent report (on the 1992 TSA), called for a continuing evaluation and—having noted that "many of the factors affecting the quality and feasibility of state NAEP are the same as those affecting national NAEP"[1]—proposed that the evaluation be expanded to include the full NAEP program. It also recommended continuing research and development in the important area of performance standards.

[1] The National Academy of Education, *The Trial State Assessment: Prospects and Realities* (Stanford, CA: Author, 1993), 104.

With the Improving America's Schools Act of 1994, Congress adopted many of the Panel's recommendations. Importantly, the legislation authorized NAEP state assessments, mandated the continuing independent review of the entire NAEP program, and directed that the state assessments and achievement levels be used on a developmental basis until the Commissioner of Education Statistics made a final determination of their validity and utility.

In this, its fourth report, the Panel presents recommendations and findings specific to its evaluation of the 1994 TSA in fourth-grade reading and offers several general conclusions regarding the state NAEP assessments. This fall, the Panel will conclude its work by releasing a capstone report that builds on these conclusions, and on its previous reports, to address issues in the redesign of NAEP for the year 2000 and beyond.

## *Dimensions of the Evaluation*

In preparing this report, the Panel has drawn upon its extensive experience with the previous TSAs as well as studies and papers commissioned specifically for the 1994 assessment. In particular, the Panel found that the guiding principles articulated in its third report to Congress remain highly relevant and continue to shape the perspective for its evaluation. These principles, revised and regrouped for clarity, are presented on pages 3 through 5 of this report.

The Panel gathered evidence regarding several dimensions of the 1994 assessment and weighed them against its principles. Some of these dimensions—content validity, sampling, assessment administration, and reporting and dissemination—have been central to the Panel's considerations of each of the previous TSAs. For 1994, the Panel also updated its conclusions regarding the National Assessment Governing Board (NAGB) achievement levels and added new emphases on scaling and analysis and on the assessment of students with disabilities or limited English proficiency.

These various dimensions are discussed in chapters 2 through 7 of the report; the Panel's primary conclusions and recommendations on each are presented below.

### *Content Validity*

The 1994 NAEP reading assessment marked the second use of the reading framework developed in 1991. A portion of the item pool was released to the public and replaced between the 1992 and 1994 assessments, but the overall parameters of the assessment were held constant, allowing reading trends to be measured for the first time on tasks that reflect current understandings of reading and reading assessment. The Panel reviewed the framework and items for content validity after each of the two assessments and, in each instance, concluded that the NAEP reading assessment was a reasonable representation of current theories in reading, a reasonably valid measure of reading achievement in the nation, and relevant to everyday classroom practice. The Panel commends NAGB and NCES for building a challenging assessment of reading achievement that extends beyond simple mastery of the mechanics of reading to include the reader's ability to draw meaning from text and to communicate this understanding to others.

Furthermore, the Panel concluded that the decision to hold frameworks in reading and other content areas constant over several assessment cycles was praiseworthy—a judgment that was confirmed by the strong interest of NAEP constituents in using 1994 results to gauge the progress of their students over time. Based on these findings, the Panel recommends that the general structure (framework) of the present reading assessment be maintained through the year 2000 or 2002.

During the evaluation, the Panel's reading experts also noted aspects of the framework and item pool that could be improved, although none of these shortcomings were sufficient to undermine the content validity of the 1994 assessment in a substantial way. More specifically, the 1994 fourth-grade assessment contained relatively few items that were within the scope of the least able students, making it difficult to get precise and reliable estimates of achievement for those at the lower end of the NAEP scale. Some unevenness of item quality was also observed. Specifically, some of the scoring guides for constructed-response items were inconsistent with other features of the items or with the directions given to students, and a number of the more difficult items failed to capture the essential features of advanced reading achievement. The Panel judges each of the above to be areas in which the NAEP contractor should begin improvements immediately, in preparation for the next NAEP reading assessment.

Finally, the Panel points out, as it has in its previous reports, that under the current funding and development process there is little time for planned, farsighted content development. In particular, during years when new frameworks are adopted, the NAEP contractor has typically had less than six months in which to develop field test materials before these materials must be finalized. Nevertheless, assessment tasks that are produced during this one brief period set the tone for all future assessments until the next revision of the framework, eight or ten years in the future.

The Panel therefore recommends that, for every NAEP subject area, NAGB and NCES adopt a process that allows new research and development to begin several years before the framework is scheduled to be revised and a new trend line begun. This research and development could progress in a relatively modest manner through successive pilot studies and small-scale trials targeted at particularly challenging research problems. When it is time to begin the actual revision, Congress, NAGB and NCES should allow for a framework and item development cycle that is substantially longer and more integrated than the current one. The Panel further recommends that Congress forward fund NAEP in order to facilitate this process.

## *Sampling and Assessment Administration*

As it had in 1990 and 1992, the Panel once again concluded that both sampling and administration for the 1994 TSA were done well, were generally consistent with best practice for major surveys of this kind, and, with the exceptions noted below, produced valid and useful state results. Two areas of concern were identified, however.

First, and most important, substantial problems were found with the samples of nonpublic schools that were—at the Panel's previous recommendation—added to the TSA for the first time in 1994. The Panel's motivation to include nonpublic schools

was based on its *inclusiveness principle,* and the Panel's intention was to aggregate the nonpublic school results with those from public schools in order to generate better overall state composites. However, NCES adopted more extensive reporting plans after determining that it would be difficult to recruit nonpublic schools without offering them separate reports of student achievement by type of school control.

Upon reviewing the evidence, the Panel concluded that the 1994 state samples of nonpublic schools were not large enough to support separate reporting. In addition, participation rates for originally-sampled nonpublic schools were unacceptably low in approximately 40 percent of the states, and final samples were biased, in many cases, by the fact that certain kinds of nonpublic schools were much less likely to participate than others.

The Panel recommends that NAGB and NCES stop separate reporting of state-level nonpublic school results but, where participation rates are sufficiently high, continue reporting state-level results for public and nonpublic schools combined and for public schools only. Furthermore, the reports should include prominent warnings about the invalidity of simplistic comparisons between public and nonpublic schools in order to discourage efforts to derive such comparisons by subtracting public school means from the combined public and nonpublic school results. These warnings should be illustrated by concrete examples to underscore their significance.

At the same time, NAGB and NCES should explore alternative strategies (other than separate state-level reporting) for motivating the participation of nonpublic schools. One proposed course of action would be to offer more detailed reporting of private school results at the national level by basing the analyses on aggregated data from the national and state samples of nonpublic schools. (For example, NAEP could break out results by more detailed categories of private schools.)

A second area of concern to the Panel involves the participation of *public* schools. Although the Panel found that in 1994, for nearly all states, the participation rates for originally-sampled public schools ranged from acceptable to good, strong indications have emerged that the data collection burden on the states, especially small states, may begin to threaten school and hence state participation rates in years when multiple subjects and grades are assessed.

The Panel recommends that NCES and NAGB consider design changes that could decrease sample size requirements or otherwise reduce burden without compromising the overall quality of the assessment. Applicable design changes could include relatively circumscribed modifications, such as applying the principles of finite sampling to create a different set of rules for the smallest states. Reduced respondent burden could also be effected as one outcome of a more radical redesign of NAEP, and various versions of the latter are currently being debated by NAGB and other interested parties.

### *The Assessment of Students with Disabilities or Limited English Proficiency*

The 1994 assessment cycle occurred at a time when NCES and NAGB were beginning to re-examine NAEP policies regarding the exclusion and assessment of students with disabilities or limited English proficiency (IEP and LEP students). In 1994, NCES

gathered data from several sources and met with representatives of the disability and bilingual communities to discuss the best methods for increasing inclusion in NAEP. The Panel, for its part, collected new data for samples of fourth-grade IEP and LEP students who had been selected for participation in the TSA, then shared its preliminary findings with NCES. The results of these various efforts led to a set of revised exclusion procedures and new allowances for accommodated assessment that were tried out in the 1995 field test and implemented, in a controlled design, in 1996.

The Panel's 1994 study indicated that school personnel in different states tended to interpret the (old) exclusion guidelines differently. Thus, on average, IEP students with the same level of ability would be included in some states and excluded in others. The Panel also found that a high proportion of IEP students (perhaps as many as 85 percent) could read well enough to participate in NAEP and be included in estimates of overall state achievement. However, the current NAEP reading test is not particularly well suited to the reading abilities of the many IEP students who are reading a grade or more below grade level. A more appropriate measure for these students would address the same reading outcomes but be based on less difficult reading passages.

The Panel's study of LEP students indicated that a significant proportion of LEP students also read well enough in English to participate for the purpose of contributing to overall state NAEP results. Disturbingly, among the LEP students sampled for the Panel's study (which was limited to LEP students who had attended English-speaking schools for at least two years), more than half had been excluded from the TSA. This was true even though more than three-quarters of the Panel's sample had been in English-speaking schools for more than four years—essentially their entire school careers. The Panel acknowledges that NAEP may not offer these students an optimal opportunity to demonstrate their competence, particularly in content areas such as mathematics or science. In reading, however, the Panel believes that it is reasonable to ask how well these students are able to read in English. Moreover, the education fortunes of LEP students may too easily drop from sight if they are excluded from major assessment efforts.

Finally, the Panel's studies found that teachers of both IEP and LEP students were likely to propose testing accommodations for high percentages of their students. Thus, when accommodations are offered, inclusion may be increased, but the overall numbers of students assessed under standard conditions may actually go down. This is problematic because scores obtained under nonstandard conditions are much more difficult to interpret.

Based on its research findings, the Panel makes the following recommendations.

1. NCES and NAGB should continue efforts to encourage greater participation of students with disabilities or limited English proficiency in the current NAEP assessments. At the same time, they should continue research to identify adaptations or accommodations for each of these groups that would provide more valid measures of subject-area achievement as specified in the NAEP content frameworks.

2. Results for students with disabilities or limited English proficiency assessed under standard conditions should be aggregated with results for all other students in producing the overall and subgroup achievement estimates normally reported for the nation and the states. The results for these populations should *not* be disaggregated or reported separately.

3. NAEP should also work to develop assessments that can measure accurately over a broader range of student proficiency levels and thereby provide better estimates at both ends of the achievement distribution. For efficiency, such an assessment would almost certainly require some adaptive mechanism (computerized or otherwise) for matching students with assessment tasks appropriate to their levels of proficiency.

## *Scaling and Analysis*

The procedures used for scaling and analysis in the TSA are generally the same as those used in the national NAEP, and analyses for the two assessments are largely interconnected. In 1994, two technical errors affecting state scores were discovered, and an unexplained but statistically significant drop in performance was observed in the national reading results at grade 12. These occurrences led the Panel to give greater attention to scaling and analysis in its evaluation of the 1994 TSA than it had in its previous evaluations.

In general, the Panel concluded that NCES and its contractors continue to make use of sophisticated methods to solve challenging measurement problems posed by recent innovations in testing and to produce generally high quality data. At the same time however, the system appears to be showing strains that allow errors to creep in, in addition to lengthening the time to reporting. Factors contributing to these strains have included pressure for frequent enhancements to the assessments, increased analysis volume, and policy pressure to reduce time to reporting.

Recent efforts to report short-term trends based on the main NAEP assessments have also uncovered some potential problems, related to the fact that many small modifications in items and procedures have been permitted between assessments. In particular, the accuracy of the 1994-to-1992 12th-grade equating may have been affected because the proportion of multiple-choice items was substantially higher in the common item pool that served as the basis for the equating than it was in the overall assessment. If the two item types indeed measure somewhat different skills, then the link was not a good proxy for the whole.

While no one major problem with the 1994 scaling or analysis was observed, the accumulation of smaller problems suggests that a modified assessment design would better fit the size and objectives of the current NAEP program. The Panel therefore supports NAGB's efforts to develop a new, more streamlined design for NAEP.

In the meantime, the Panel makes the following recommendations to help ensure the integrity of NAEP results:

1. Any significant change in performance on the short-term trends should routinely be checked for reasonableness against other sources of trend data—sources such as the long-term NAEP trend data and state assessment trend data—before the results of the short-term trend are reported.

2. NCES should conduct or commission additional studies to validate the current analysis and scaling models. These studies should include research on the strength of the models being employed and the robustness to violations of assumptions.

3. Additional procedures designed to verify the integrity of the NAEP data prior to its release should be investigated, and NCES should continue to give priority to the timely release of high quality technical reports that provide thorough documentation of all design related, technical, and psychometric activities associated with the assessments.

## *Achievement Levels*

The achievement levels established by NAGB for the 1992 reading assessment were again used for reporting the 1994 assessment. The Panel realizes that reporting by performance standards is greatly valued by much of the NAEP (and TSA) constituency. Nevertheless, the Panel continues to question the reliability and validity of the current achievement levels. At the time of its evaluation of the 1992 achievement levels, the Panel concluded that 1) the standard-setting method had led to serious internal inconsistencies that could have especially troubling consequences if the mix of item types changed over time and 2) the distributions of student performance established by the achievement-level cutscores was not reasonable based on comparison to the distributions suggested by various non-NAEP measures. In particular, the weight of the evidence suggested that the 1992 achievement levels were set too high.

Although the achievement-levels contractor fielded a study in 1994 that putatively addressed the second of these concerns, the Panel concluded that the design of the study did not permit confirmation of specific cutscores. The study was therefore not particularly informative with respect to the Panel's conclusion that the cutscores had probably been set too high.

The Panel also examined the results for the 1994 U.S. history and world geography achievement levels in order to determine whether they would exhibit better internal consistency or a better match to external criteria than had the 1992 reading and mathematics achievement levels.[2] In fact, the Panel once again found troubling differences in achievement-level cutscores set using dichotomous versus partial-credit (extended-response) items. Although not as dramatic as the differences found for the 1992 achievement-levels, the 1994 results again showed that levels set using extended-response items were considerably higher than those set using multiple-choice or dichotomously-scored constructed-response items.

The Panel also examined the achievement levels in relation to performance on the AP examination in U.S. history.[3] Many colleges and universities give college credit for AP courses taken in high school if students score three or better, and the Panel found that 2.8 percent of the country's high school seniors met this criterion on the AP U.S. history examination in 1994. By contrast, NAEP classified only 1 percent of high school seniors at the advanced level in this subject. Moreover, the percentage passing

[2] U.S. history and world geography were assessed nationally in 1994 but were not included in the TSA. It was not therefore in the Panel's purview to conduct a formal evaluation of the achievement levels set for these subjects. However, to the extent that the data were readily available, the Panel believed it should determine whether or not the results from these new level-setting efforts confirmed the Panel's earlier findings.

[3] Only U.S. history could be considered because no AP examination is offered in world geography.

the AP criterion would have been even higher if AP programs had been available in all U.S. high schools instead of only half of them. These findings provide additional evidence that the Governing Board's achievement levels are set too high, that is, that the achievement levels identify fewer 12th graders as advanced than actually are performing at an advanced level.

Based on its accumulated evidence concerning the achievement levels and the process by which they were set, the Panel makes the following recommendations.

1. NAGB should institute a competition for the design of new methods for setting performance standards for all NAEP subjects with the goal of having a new method in place by the time of the year 2000 NAEP assessment.

2. In the interim, current achievement levels should be accompanied by a warning stating that results should be interpreted as suggestive rather than definitive because they are based on a methodology that earlier evaluation panels have questioned in terms of accuracy and validity.

### *Reporting and Dissemination*

In considering the quality of NAEP reporting since the inception of the TSA, the Panel has identified four criteria fundamental to successful reporting:

- The accuracy of the results;
- The likelihood that the results will be interpreted correctly by the intended audience;
- The extent to which the results are accessible and adequately disseminated; and
- The timeliness with which the results are made available.

With respect to the first three criteria, NCES, NAGB, and the NAEP contractors have made steady progress. For example, innovative graphic formats intended to convey the statistical significance or insignificance of differences between states and across time have been tried after each TSA, and the map graphics introduced in 1992 proved more successful than earlier efforts. The 1994 reports retained most of these graphics and also addressed other concerns about the interpretability and accessibility of the results by introducing more charts, visually simplifying the data tables, using more white space, and generally shortening the reports. NCES has also begun a series of focused reports that highlight specific findings from each assessment cycle, and these also have been well received. Two such reports have been scheduled for the 1994 reading assessment. Further, in response to the expressed need of the state assessment directors for a brief and readable summary of results that they could distribute to educators and policy makers in their states, NCES produced a four-page brochure that was released with the main reading reports in March 1996.

The 1994 TSA was particularly problematic, however, with respect to timeliness. Despite efforts by NCES, NAGB, and the NAEP contractors to speed up reporting, the new *First Look* report, which contained only summary findings for the 1994 reading assessment, was not released until April, 1995 (13 months after the administration). The main reading reports did not appear for nearly another year after that—the longest lag between assessment and reporting that has occurred to date. Factors which contributed to the delay included unexpected data problems, shifting program priorities, and competition for the services of qualified analysis staff. The Panel strongly encourages NCES and NAGB to continue to press for quicker and more timely reporting while also being careful to maintain the quality and integrity of the data.

## *Utility of the TSA*

The final perspective that bears on the overall evaluation of the TSA, and in effect subsumes all other perspectives, concerns its utility. As suggested above, utility must rest, firstly, on the validity and reliability of the data. Beyond this, the results must be timely, accessible, and policy relevant, and the program must be perceived as useful and valuable by the major customers of the information it provides—particularly the states. To investigate the latter, the Panel commissioned surveys and case studies of NAEP's perceived influence after the release of each round of TSA data, concluding with a set of case studies and a mail survey of state assessment directors, mathematics specialists, and reading specialists in December, 1995. Throughout its evaluation, the Panel also monitored media coverage of NAEP and the TSAs and followed the opinions and actions of other NAEP stakeholders.

### *Utility of NAEP Data to the States*

For the most part, the Panel concluded from these efforts that state NAEP has become a valued indicator of educational progress and has served particularly to provide an independent validity check on the states' own assessments. In Rhode Island, for example, the state reading specialist reported that the 1994 TSA reading results provided important evidence for the success of an ongoing reading initiative. The external monitor role has been especially important during a period when many state assessments have undergone radical reform, making upward or downward trends in their results particularly difficult to interpret.

Several factors contribute to state NAEP's credibility and hence its value to the states as an external monitor. These include the assessment's forward-looking content and format, the secure status of its testing materials, and the rigorous statistical standards maintained in data collection, analysis, and reporting. (The long lag time to reporting, and the lack of a stable assessment schedule against which states can plan, however, are two factors cited repeatedly by the states as diminishing state NAEP's utility.)

When state NAEP results have yielded dramatic or unexpected results, particularly when a state's students performed worse than expected, considerable public debate has followed. North Carolina and California both provide notable, and very different, examples of this effect.

In 1990, North Carolina educators were dismayed to discover that the state's students had done much worse on the NAEP mathematics assessment than the educators had expected, based on results from the state's own assessment, a commercially available, norm referenced test. During the subsequent debate and discussion, decision makers concluded that North Carolina teachers generally lacked certain key understandings that were required to implement their recently introduced, forward-looking mathematics curriculum successfully. Information from the NAEP background questions on instructional practices helped North Carolina reach this conclusion, and the state subsequently undertook an intensive in-service training program that was based in part on materials and data from the 1990 NAEP. These remediation efforts appeared to be successful in that the 1992 TSA showed a significant gain in eighth-grade mathematics achievement for the state.

In California, educators and the public were also shocked when fourth-grade reading achievement estimates from the 1994 TSA showed California performing significantly worse than it had done in 1992, and positioned virtually at the bottom of the distribution of participating states. This information was particularly important in view of the fact that California's own assessment system has been in disarray for the past several years, precluding any meaningful assessment of performance trends from that particular source. However, in the resultant furor, most commentators simply pointed to the TSA results as further evidence for what they already felt was wrong with the state's education system—whether that was crowded classrooms or the state's whole language reading curriculum.

Besides using NAEP as an external monitor of achievement, about 60 percent of the states that undertook revisions to their mathematics or reading curricula during the past five years reported NAEP as a notable source of ideas. Similar numbers referred to NAEP as a model, or a source of external validation, for changes to their reading or mathematics assessments. State educators, for example, have closely followed NAEP's pioneering efforts to set performance standards, and both assessment directors and curriculum support staff have used NAEP's external credibility to argue for such desired objectives as better alignment with National Council of Teachers of Mathematics (NCTM) standards or with reading standards based on reading for meaning, higher order skills, and real-world reading tasks.

## *Contributions to the National Debate*

Interestingly, state NAEP has broadened NAEP's influence not only at the state level, as might be expected, but also at the national level. NAEP has been adopted by the National Goals Panel as the primary indicator of progress towards goal three, which states that "By the year 2000, American students will leave grades four, eight, and twelve having demonstrated competency in challenging subject matter..."[4]

NAEP also routinely receives national press coverage after each of its major data releases. The coverage has tended to be more widespread when regional media are able to report on results for their own states as well as for the nation. Publications

[4] National Education Goals Panel, *The National Education Goals Report: Building a Nation of Learners* (Washington, D.C.: Author, 1991), 10.

devoted to education news, such as *Education Week*, also contain frequent references to NAEP, both as a unique source of information about education achievement and as a model for current assessment practices.

## *The Impact of State on National NAEP*

When state NAEP was authorized by Congress on a trial basis in 1988, one of Congress' central concerns was whether state NAEP would have a deleterious effect on national NAEP. By asking this question, Congress was tacitly affirming the importance of protecting the integrity of national NAEP and expressing a concern that state NAEP might have a negative impact on state participation in national NAEP, especially in the case of small states. There is little evidence that this has happened to date.

Rather, the Panel believes that an implicit, mostly unspoken *quid pro quo* has developed between the states and NAGB, by means of which the states are willing to participate in national NAEP *at least in part* because of the value they get from participation in state NAEP. Since 1990, the Panel has observed movement from guarded cooperation among participating states to general anticipation when state NAEP results are about to be released. Positive attitudes toward state NAEP can only grow if NCES and NAGB are successful in addressing the relatively few persistent concerns, such as the uncertainty of the assessment schedule, that states have cited repeatedly. As a result, the Panel suggests that, in the unlikely event that Congress were to recommend the abandonment of state NAEP, the motivation of the states to continue in national NAEP could drop precipitously.

As a result, in contrast to its original conclusion at the end of the evaluation of the 1990 TSA, which was simply that state NAEP had had no deleterious effect on national NAEP, the Panel now believes that the future of national NAEP has become intertwined with the future of state NAEP. State NAEP has greatly increased the visibility and perceived utility of the entire NAEP program, and suggestions for merging the state and national samples continue to arise (although it is not evident that such a merger would be feasible or significantly reduce burden).

There is obviously also an interaction between monies spent on state NAEP and monies available to maintain a quality national NAEP program, but the nature of this interaction is complex. On the one hand, the substantial funds spent for state NAEP cannot then be spent for other NAEP activities. On the other hand, the heightened visibility conferred by state NAEP may result in a net increase in national NAEP resources. For example, the substantial framework and item development efforts that have characterized the last several years have benefited both programs and almost certainly would not have been funded without the impetus of state NAEP.

## *The Panel's Recommendation for the Continuation of State NAEP*

Based on its evaluation of the TSAs, the Panel concludes that state NAEP has been shown to be a valid, reliable, and useful measure of student achievement, and that it aligns favorably with the Panel's quality, utility, and state indicator principles. For

these reasons, the Panel recommends that state NAEP be continued, and that it be moved from developmental to permanent status when NAEP is next reauthorized. However, in light of its size and cost, the Panel further recommends that the scope and function of state NAEP be reviewed regularly, and particularly after any substantial change in mission or design. Such re-evaluation should be done in the context of the overall NAEP program and with the abiding aim of providing the best and most useful information about student achievement for the nation.

There are areas, however, in which it is not yet possible to determine the course that will best serve NAEP's mission and goals. Some of these areas, which should continue to be examined in the near future, include

- The viability of continuing to assess nonpublic schools in the state NAEP program;
- The value and feasibility of grade-12 state assessments;
- The tension between including as many students with disabilities and limited English reading skills in the assessment as possible, and the cost of doing so;
- The adequacy of the present NAEP design to meet the increasing demands of NAEP's stakeholders while still satisfying the Panel's *quality principle;* and
- The development of improved performance standards for reporting NAEP results.

In the fall, 1996, the Panel will present its capstone report. Building upon the Panel's previous work, the report will look forward to the year 2000 and beyond, considering recommendations for the design of a NAEP program that offers quality assessments for the nation and the states and also anticipates the changing nature of education practice as the latter will be influenced by technology and by our developing knowledge of learning and human cognition.

# 1 Introduction

## The Context for the Panel's Evaluation of the 1994 TSA

Since its inception in 1969, the National Assessment of Educational Progress (NAEP) has been the nation's leading indicator of what American students know and can do. The high technical quality of the assessment and its independence from education and political pressures have enabled NAEP to reliably monitor changes in education achievement and practices for nearly three decades. Moreover, as the only education assessment administered to a representative sample of American students, NAEP has been able to track changes not only for the population as a whole, but for important subgroups as well. For example, analyses of NAEP trend data in the past decades revealed a significant narrowing of the achievement gap between African American and white students during the late 1970s and 1980s, while subsequent changes in these same data patterns alerted observers to an apparent reopening of that gap in the 1990s. Similarly, NAEP has pointed to important trends in specific subject areas, documenting, for example, the declining achievement of American students in science between 1970 and 1990, as well as the limited time spent on science instruction in most elementary classrooms. These data have provided evidential support for groups such as the American Association for the Advancement of Science and the National Science Foundation, who have argued for greater attention to science education in American schools. Finally, NAEP data were also cited in discussions and debates that led the Governors and then Congress to target improved science achievement as an important national education goal.

The National Assessment of Educational Progress has thus been a long standing and useful source of information for national education reformers and policy makers. It was not until 1988, however, that NAEP could begin to play a similar role for the individual states. Prior to the passage of Public Law 100-297 in that year, NAEP was prevented both by its design and by its mandate from collecting and reporting data at the state level. That prohibition was lifted by Congress in 1988 with its authorization of an experimental new component to the NAEP program—the Trial State Assessments (TSA). This action reflected both the educational and the political context of the times. Specifically, by approving and funding the state-by-state assessment, Congress was responding to increased education reform activity in the states and to expanded interest in using state-level data to monitor progress. At the same time, by making participation in the TSA voluntary, Congress demonstrated its continuing respect for the constitutional authority granted to the states for the education of their residents. Finally, by authorizing state NAEP on a "trial" basis only, Congress showed recognition that such a massive expansion of the NAEP program was unproven in its direct effectiveness and in its overall impact. Congress further underscored this recognition by mandating that the trials be independently evaluated so as to determine whether they provided valid, reliable, and useful data for the states.

The 1994 state NAEP assessment in reading brings to a close the series of TSAs authorized under the original 1988 legislation and subsequently extended through 1994. With the conclusion of the TSAs comes also the conclusion of the evaluation of the program under the auspices of The National Academy of Education (NAE) Panel

on the Evaluation of the NAEP Trial State Assessment. The purpose of this report is twofold: first, to present the Panel's specific findings and recommendations stemming from its evaluation of the 1994 TSA in fourth-grade reading and second, to offer several general conclusions regarding the larger TSA program. In a subsequent capstone report, the Panel will build on these conclusions to provide analysis and recommendations for the entire NAEP program in the year 2000 and beyond.

## *History of NAEP TSA Evaluations*

Over the past six years, the NAE Panel has seen the NAEP state assessment program grow, albeit somewhat erratically, and become a highly valued feature of the NAEP program. In the face of budgetary constraints however, the final 1994 trial, like the first, included only one subject at one grade. The largest trial, in 1992, covered two subjects at grade 4 and one at grade 8. A trial at the 12th grade was never carried out. (See table 1.1.)

**Table 1.1.** ***Grades and subjects assessed in the NAEP trial state assessments 1990-1994***

| | 1990 | 1992 | 1994 |
|---|---|---|---|
| Grade 4 | ___ | Reading | Reading |
| | ___ | Math | ___ |
| Grade 8 | ___ | ___ | ___ |
| | Math | Math | ___ |

Participation by states and other jurisdictions has increased with each assessment cycle, reaching a new high of 46 participating jurisdictions with the just completed 1996 state assessment, which was also the first state assessment that was not labeled as a "trial." Additionally, state trend results, which have been available from every assessment since 1992, have helped to sustain the attention of user groups, who continue to show strong interest in the results.

Throughout this period, the NAE Panel has been involved in an ongoing evaluation of the TSA program. Three reports were submitted to Congress over the course of the first two TSAs. In them, the Panel concluded that the TSAs were successful and should be continued. Certain areas for improvement or further study were also noted however, including areas in which the full consequences could not be ascertained before the trials were scheduled to end. Furthermore, the Panel noted in the last of these reports, *The Trial State Assessment: Prospects and Realities,* that "many of the factors affecting the quality and feasibility of state NAEP are the same as those affecting national NAEP."[1] Accordingly, the Panel called for continuing evaluation of the full NAEP program, including the state assessment component. It also called for continuing research and development in the important area of performance standards.

[1]The National Academy of Education, *The Trial State Assessment: Prospects and Realities* (Stanford, CA: Author, 1993), 104.

Congress accepted many of the Panel's summative judgements when it passed the NAEP reauthorization as part of the 1994 Improving America's Schools Act. Specifically, the legislation authorized state assessments, called for the continuing independent review of the entire NAEP program—including the national assessment, state assessments, and student performance levels—and further directed that both the state assessments and the achievement levels be used on a developmental basis until the Department of Education's commissioner of education statistics determines that each is valid and useful. Many other recommendations made by the Panel were also adopted by NAEP, including the addition of private schools to the state samples, increased data collection for participating Individual Education Plan (IEP) students and Limited English Proficient (LEP) students, the creation of focused reports for specific NAEP topics, and the establishment of standing subject-matter panels to provide continuity through each stage of the assessment process. Most of these changes have contributed to recognizable improvements in NAEP; at least one, however (the recommendation to include private schools), has also had unexpected consequences, underscoring the highly complex and dynamic nature of the program and the context in which it operates.[2]

## *Guiding Principles*

Before delving into the specifics of the 1994 reading assessment, the Panel notes that this final cycle of the TSA evaluation uses and builds on its previous evaluations. This is true not only with respect to the critical questions addressed, but also with respect to the underlying perspective the Panel brings to its deliberations. That perspective is embodied in a set of guiding principles articulated by the Panel in its third report to Congress.

In preparation for the present evaluation, and for its final capstone report, the Panel has reviewed those principles, revising and regrouping them for greater clarity. Below, we restate the principles in their revised form with brief explanatory comments as appropriate.[3]

## *NAEP's Mission as an Independent Indicator of Student Achievement*

The first three principles focus on NAEP's mission as an indicator of what students know and can do. As discussed, NAEP has performed this mission for the nation as a whole for close to three decades, not only recording student achievement in key subject areas in specific years, but also monitoring changes in that achievement over the course of time. Since 1990, NAEP's mission has broadened to make available for the states the same high quality of data NAEP has long provided for the nation.

[2]See chapter 3 of this report.

[3]See The National Academy of Education, *The Trial States Assessment: Prospects and Realities* (Stanford, CA: Author, 1993), 93-98, for the earlier version and more expanded discussion of these principles.

Moreover, in this era of global competition and education reform, both the states and the nation have developed heightened interest in comparing the performance of their students and of their education systems with those of other countries. Of all the large-scale assessments in the United States, NAEP is in the best position to be linked in a meaningful way with similar assessments internationally.

*The National Indicator Principle*

NAEP should continue to be the key independent indicator of what the nation's students know and can do, providing trend data on student academic performance in key subject areas.

*The State Indicator Principle*

Because states have constitutional authority for education, NAEP should continue to play an additional, needed role by providing independent information to the states on the educational progress of their students.

*The International Indicator Principle*

The United States must compare its education practices and results with those of other nations and, where possible, learn from the education practices of others. NAEP should therefore, to the extent possible, be linked with major international assessments.

## *NAEP's Fundamental Criteria of Excellence: Quality and Utility*

Recent review of the Panel's guiding principles has led to the addition of two new principles that reflect the very basis on which the success of the NAEP program must be judged. Previously, the Panel had taken these criteria of excellence for granted, using them in deliberations but failing to articulate them explicitly as part of the fundamental perspective guiding its work. These two criteria of *quality* and *utility* provide the goal and the rationale for many of the more specific principles and recommendations this Panel has adopted over the past six years.

*The Quality Principle*

NAEP should be exemplary in the development and use of assessment and reporting techniques and practices that produce reliable, fair, and valid estimates of student achievement.

*The Utility Principle*

The NAEP data and program must be useful for a variety of stakeholders, including policy makers, the public, educators, and researchers.

If *quality* and *utility* are the goals, what specific characteristics should NAEP promote to achieve those goals? Below, the Panel articulates five enabling characteristics that it believes will help foster a high quality and useful assessment. As is the case with most such sets of principles, those that follow are purposefully written so as to allow flexibility in interpretation. They are not stated in absolutes because the relative emphasis placed on each must be determined by the extent that it, in conjunction with the other principles, contributes to the overall quality and utility of the entire NAEP program.

*The Comprehensiveness Principle*

NAEP frameworks and content must be comprehensive, reflecting both the range of current education practice, sound theory, and research about human learning and performance.

*The Relative Stability Principle*

NAEP frameworks must be in place long enough and assessed regularly enough to establish meaningful trend lines. At the same time, the framework revision process must remain attuned to the natural evolution that occurs in subject-matter fields.

*The Inclusiveness Principle*

To the degree technically, ethically, and financially possible, NAEP should assess an inclusive sample of all children in the designated age or grade populations.

*The Policy Relevance Principle*

NAEP must collect data relevant for policy makers and education decision makers, and report the data in a timely fashion, while maintaining its integrity uncorrupted by political pressures.

*The Public Information Principle*

NAEP data and reports must be accurate in content, comprehensible in format, and readily accessible to all relevant stakeholders.

Throughout the chapters of this report, as well as the capstone document that will follow, the Panel uses the principles outlined above to guide its evaluation and its recommendations for the future of NAEP. Where appropriate, these reports refer to specific principles and to the interplay among them as they affect the Panel's deliberations.

## The Panel's Forthcoming Capstone Report on the Future of NAEP

As noted, whereas this report deals with the specifics of the 1994 TSA, the Panel will also be issuing a capstone report that summarizes the lessons learned from the three TSAs and looks to the future of NAEP as we approach the 21st century. The past six years since the initiation of the Trial State Assessments have witnessed substantial shifts in the scope and purposes of the overall NAEP program (e.g., size, audience, link to accountability, and goals). Indeed, the TSA itself has played a prominent role in the evolution of the overall NAEP program. The capstone report will therefore consider the extent and implications of the interlinkages that currently exist, and those that might exist between state and national NAEP. More broadly, the capstone report will present an overview of the major issues and choices that confront NAEP in the closing years of the 20th century and offer the Panel's perspective on them. This future-looking capstone report will come at a particularly opportune time, since the National Assessment Governing Board (NAGB) is currently considering major revisions in the NAEP design, expected to take effect with the 1998 reauthorization of NAEP.

## The Structure of this Report

The 1994 evaluation incorporates each of the topics addressed in the Panel's previous evaluations of the 1990 and 1992 TSAs. In addition, the evaluation introduces some new areas of investigation, including the assessment and exclusion of students with disabilities or limited English proficiency, and the extensive scaling and analytic activities involved in producing NAEP results. The report is organized into the following chapters:

***Chapter 2. The Content Validity of the 1994 Reading Assessment***

Chapter 2 examines the content validity of the 1994 reading assessment. While the 1994 reading assessment reflected the same content framework as that used in 1992, approximately one-quarter of the reading tasks were new. The chapter begins by reviewing the overall structure of the 1994 reading assessment. This is followed by a brief description of the Panel's empirical studies and main findings regarding the content validity of the reading assessment. In the section that describes the Panel's main findings, several paragraphs are devoted to a discussion about the problematic areas identified in the assessment and in the framework. The chapter concludes by providing a set of recommendations for improving the assessment development process in general, and the reading assessment in particular.

***Chapter 3. Sampling and Assessment Administration for the 1994 TSA***

Organized into three main sections, chapter 3 focuses on the sampling and assessment administration for the 1994 TSA. The first section examines issues related to the sampling and participation of public schools and public school students, and the second section addresses the same topics for nonpublic schools. The last section presents information about the administration of the TSA and includes specific

findings from the Panel's study. The topics covered include comparisons of TSA and national performance, comparisons of monitored and unmonitored TSA sessions, and the relationship between student performance and characteristics of the session administration.

***Chapter 4. The Assessment of Students with Disabilities or Limited English Proficiency***

Chapter 4 examines the assessment and exclusion of these two groups of students with special needs. The chapter begins by providing background information on the IEP and LEP exclusion procedures in effect through 1994 and the exclusion rates observed in each of the three TSAs. Following this section, there is a description of included and excluded IEP and LEP populations based on data from the IEP/LEP Student Questionnaires that were administered by NAEP for the first time in 1994. Next, the research questions, methods, and findings of the Panel's study of IEP students in the 1994 TSA are presented, followed by comparable data from the Panel's LEP study. The chapter ends with a brief discussion of the procedural changes for IEP and LEP students that have been introduced by NAEP since 1994, followed by the Panel's recommendations regarding appropriate treatment of these populations in future assessments.

***Chapter 5. Scaling and Analysis of the 1994 Reading Assessment***

Chapter 5 begins with an overview of the analytic procedures used by NAEP to produce summary measures of student achievement. The subsequent sections summarize major innovations that have been implemented since 1990, discuss the potential for system errors highlighted by two events that occurred during the analysis of the 1994 reading assessment, and offer recommendations for improving NAEP's scaling and analysis procedures in the future.

***Chapter 6. Reading Achievement Levels***

Chapter 6 begins with a brief overview of the process by which the reading achievement levels were set in 1992. It then summarizes the work the Panel undertook to evaluate the achievement levels and reviews a study commissioned by NAGB to respond to the Panel's criticisms. The achievement-level results from the 1994 U.S. history and world geography assessments also receive some attention because, although not directly applicable to the TSA, they illustrate several key problems that were evident with the achievement levels set in 1992 and that are not yet resolved. The chapter concludes with a Panel recommendation that addresses the continued use of the current achievement levels.

***Chapter 7. Reporting and Dissemination for the 1994 Reading Assessment***

Chapter 7 is organized around four criteria that the Panel deems fundamental to successful reporting: the accuracy of the results, the likelihood that the results will be interpreted correctly by the intended audience, the timeliness with which the results are made available, and the extent to which the results are accessible and adequately disseminated. Suggestions for increasing the accessibility of NAEP data are provided at the end of the chapter.

***Chapter 8. Conclusions and Recommendations***

In chapter 8, the Panel presents conclusions and recommendations that are grounded in both the empirical research commissioned by the Panel and the Panel's guiding principles. The chapter begins with brief summaries of the Panel's findings for each of the major issues discussed in the report, then continues with a discussion of the utility of the TSA, the impact of state NAEP on national NAEP, and the Panel's recommendation for the continuation of state NAEP. The chapter concludes by raising some issues for continued examination in the future.

# 2 The Content Validity of the 1994 Reading Assessment

## Introduction

During the 1990s, NAGB substantially revised the content of the NAEP assessments in an effort to provide more valid measures of students' academic achievement in each of the targeted subject areas. The new assessments reflect contemporary research about teaching and learning and also the increasing sophistication of assessment technology. With regard to reading, the subject assessed in the 1994 TSA, these innovating influences are manifest in such features as

- A research-based model of how students read for meaning;
- An emphasis on reading outcomes rather than isolated skills; and
- Reading passages and tasks that are typical of real, everyday reading.

When the Panel began its evaluation of the 1994 TSA, one critical component of the work was to consider whether or not this new reading assessment had, as hoped, provided a meaningful, appropriate, and useful measure of reading achievement for the nation and the states. The Panel had already considered this question once, in conjunction with its evaluation of the 1992 TSA. At that time, the Panel concluded that the NAEP reading assessment "incorporated significant positive advances including a framework consistent with current theory and practice, interesting and authentic passages, longer testing time per passage, and a high proportion of open-ended items."[1]

Although the overall structure of the 1994 reading assessment was the same as the structure used in 1992, approximately one-quarter of the reading tasks were new. For this reason, the Panel determined that a reconsideration of content validity was warranted. In addition, the Panel brought to its 1994 evaluation an enriched perspective informed by new advances in the theoretical understanding of reading and by a growing body of practical experience with new-style reading assessments.

Overall, the Panel's new deliberations affirmed its previous conclusions about the validity of the NAEP reading assessment. Additionally, a number of new insights emerged that have allowed the Panel to advance a more coherent set of recommendations for the improvement of future reading assessments.

In the chapter that follows, the Panel begins with an overview of the reading assessment. This is followed by a brief description of the Panel's empirical studies and main findings regarding the validity of the 1994 reading assessment. The chapter concludes with a set of recommendations for improving the assessment development process in general and the reading assessment in particular.

[1]The National Academy of Education, *The Trial State Assessment: Prospects and Realities* (Stanford, CA: Author, 1993), 68.

## *The Overall Structure of the Reading Assessment*

The NAEP reading assessment uses a model of reading achievement that extends beyond simple mastery of the mechanics of reading to include the reader's ability to understand what he or she has read, to extend and elaborate this understanding by drawing upon other things already known, and to critically evaluate the meaning and utility of each new text in light of this broader understanding. In addition, the model assumes that the reader's capacity to draw meaning from the text also involves the ability to communicate that understanding to others. The assessment therefore requires students to read fairly lengthy texts drawn from actual books and stories they might encounter in class and to demonstrate their understanding by responding to a series of questions, some of which the student must answer in his or her own words. The latter are referred to as constructed-response questions.

Consistent with other NAEP assessments, the 1994 reading assessment was not designed to generate scores for individual students. Rather, different students received different subsets of tasks, and analysts combined information across students and tasks to provide group estimates of reading achievement. In 1994, the fourth-grade item pool consisted of 84 questions or tasks, organized into eight "blocks." Just over half of the reading blocks were based on stories or poems; the remainder assessed the students' ability to understand and use information in various kinds of nonfictional texts. Each student worked through two 25-minute blocks or about one-quarter of the item pool.

Figure 2.1 presents the full text of the reading passage used in one 1994 item block and four examples of the kinds of questions or tasks that accompanied it. Scoring guides for two of the questions are given in appendix A. The text, "Hungry Spider and the Turtle," is a story based on a West African folktale.

**Figure 2.1.** ***Sample reading task from the 1994 TSA***

NOTE: In the 1994 reading assessment, fourth-grade students were given 25 minutes to read this story and answer 11 questions, including five constructed-response questions. The latter required the students to write their answers in their own words.

**HUNGRY SPIDER AND THE TURTLE**
by Harold Courlander and George Herzog

Spider was a hungry one, he always wanted to eat. Everybody in Ashanti knew about his appetite. He was greedy, too, and always wanted more than his share of things. So people steered clear of Spider.

But one day a stranger came to Spider's habitation out in the back country. His name was Turtle. Turtle was a long way from his home. He had been walking all day in the hot sun, and he was tired and hungry. So Spider had to invite Turtle into his house and offer him something to eat. He hated to do it, but if he didn't extend his hospitality to a tired traveler it would get around the countryside and people would soon be talking about Spider behind his back.

So he said to Turtle:

"There is water at the spring for you to wash your feet in. Follow the trail and you'll get there. I'll get dinner ready."

Turtle turned and waddled down to the spring with a gourd bowl as fast as he could. He dipped some water from the spring and carefully washed his feet in it. Then he waddled back up to the house. But the trail was dusty. By the time Turtle got back to the house his feet were covered with dirt again.

Spider had the food all set out. It was steaming, and the smell of it made Turtle's mouth water. He hadn't eaten since sunrise. Spider looked disapprovingly at Turtle's feet.

"Your feet are awfully dirty," he said. "Don't you think you ought to wash them before you start to eat?"

Turtle looked at his feet. He was ashamed, they were so dirty. So he turned around and waddled as fast as he could down to the spring again. He dipped some water out of the spring with the gourd bowl and carefully washed himself. Then he scurried as fast as he could back to the house. But it takes a turtle a while to get anywhere. When he came into the house Spider was already eating.

"Excellent meal, isn't it?" Spider said. He looked at Turtle's feet with disapproval. "Hm, aren't you going to wash yourself?"

Turtle looked down at his feet. In his hurry to get back he had stirred up a lot of dust, and his feet were covered with it again.

"I washed them," he said. "I washed them twice. It's your dusty trail that does it." "Oh," Spider said, "so you are abusing my house now!" He took a big mouthful of food and chewed it up, looking very hurt.

"No," Turtle said, sniffing the food, "I was just explaining."

"Well, run along and wash up so we can get on with eating," Spider said.

Turtle looked. The food was already half gone and Spider was eating as fast as he could.

Turtle spun around and hurried down to the spring. He dipped up some water in the gourd bowl and splashed it over his feet. Then he scrambled back to the house. This time he didn't go on the trail, though, but on the grass and through the bushes. It took him a little longer, but he didn't get dust all over his feet. When he got to the house he found Spider licking his lips.

**Figure 2.1.** ***Sample reading task from the 1994 TSA continued***

"Ah, what a fine meal we had!" Spider said.

Turtle looked in the dish. Everything was gone. Even the smell was gone. Turtle was very hungry. But he said nothing. He smiled.

"Yes, it was very good," he said. "You are certainly good to travelers in your village. If you are ever in my country you may be assured of a welcome."

"It's nothing," Spider said. "Nothing at all."

Turtle went away. He didn't tell other people about the affair at Spider's house. He was very quiet about his experience there.

But one day many months later Spider was a long distance from home and he found himself in Turtle's country. He found Turtle on the shore of the lake getting a sunbath.

"Ah, friend Spider, you are far from your village," Turtle said. "Will you have something to eat with me?"

"Yes, that is the way it is when a person if far from home—generosity meets generosity," Spider said hungrily.

"Wait here on the shore and I'll go below and prepare the food," Turtle said. He slipped into the water and went down to the bottom of the lake. When he got there he set out the food to eat. Then he came to the top of the water and said to Spider, who was sitting impatiently on the shore, "All right, everything is ready. Let's go down and eat." He put his head under water and swam down.

Spider was famished. He jumped into the water to follow Turtle. But Spider was very light. He floated. He splashed and splashed, kicked and kicked, but he stayed right there on top of the water. For a long time he tried to get down where Turtle was eating, but nothing happened.

After a while Turtle came up, licking his lips.

"What's the matter, aren't you hungry?" he said. "The food is very good. Better hurry." And he went down again.

Spider made one more desperate try, but he just floated. Then he had an idea. He went back to the shore, picked up pebbles and put them in the pockets of his jacket. He put so many pebbles in his pockets that he became very heavy. He was so heavy he could barely walk. Then he jumped into the water again, and this time he sank to the bottom, where Turtle was eating. The food was half gone. Spider was very hungry. He was just reaching for the food when Turtle said politely:

"Excuse me, my friend. In my country we never eat with our jackets on. Take off your jacket so that we can get down to business."

Turtle took a great mouthful of food and started chewing. In a few minutes there wouldn't be anything left. Spider was aching all over with hunger. Turtle took another mouthful. So Spider wriggled out of his coat and grabbed at the food. But without the pebbles he was so light again that he popped right up to the top of the water.

People always say that one good meal deserves another.

Harold Courlander: "Hungry Spider and the Turtle," *The Cow-Tail Switch & Other West African Stories* (New York: Henry Holt and Company, Inc., 1947).

***Sample Questions***

*Multiple Choice:*

1 Why did Spider invite Turtle to share his food?
   a. To amuse himself
   b. To be kind and helpful
   c. To have company at dinner
   d. To appear generous

2. The final sentence of the story is: "People always say that one good meal deserves another." The author uses this sentence as a way of saying that
   a. Turtle and Spider both were good cooks
   b. Turtle should not have invited Spider to dinner
   c. Spider earned what Turtle did to him
   d. Spider should have cooked what Turtle liked to eat

*Constructed Response:*

3. Do you think Turtle should have done what he did to Spider? Explain why or why not?

4. Think about Spider and Turtle in the story. Pick someone you know, have read about, or have seen in the movies or on television and explain how that person is like either Spider or Turtle.

NOTE: The full scoring guides for the constructed-response items, along with examples of student responses, are shown in appendix A.

Underlying the reading assessment is a written framework that sets out the general structure of the assessment and the model of reading achievement on which it is based. The current framework, which was developed through a consensus process involving subject-matter specialists, educators, and representatives of the general public, was written prior to the 1992 reading assessment.

Using the reading framework as a guide, the NAEP contractor produced a pool of items, or tasks, for the 1992 assessment.[2] A panel of subject-matter experts, some of whom participated in the framework development process, reviewed the items and revised or eliminated those deemed problematic. Between 1992 and 1994, about one-quarter of the tasks in the reading item pool were released for public review and replaced with new tasks. The remainder of the 1992 reading tasks were reused in 1994. This continuity of framework and tasks allows NAEP to measure progress over time.

[2] The same framework and items are used for both the national and state NAEP assessments. However, national NAEP spans grades 4, 8, and 12, whereas the 1992 and 1994 TSAs in reading were limited to grade 4.

Among other things, the reading framework established an organizing structure of *situations* and *stances* intended to facilitate the interpretation of reading achievement results. By stipulating the proportion of items to be devoted to each situation and each stance at each grade level, the framework also helped ensure that the initial assessment would exhibit the emphases intended by the framework developers and that the mix of reading tasks would remain stable across assessment years, even though specific item blocks were added, and others deleted from the item pool.

The situation dimension refers to the reader's purpose for reading, which is typically associated with particular types of text and reading tasks. For example, the framework explains that "reading for literary experience usually involves the reading of novels, short stories, poems, plays and essays. In these reading situations, readers explore the human condition and consider interplays among events, emotions and possibilities."[3]

The current framework identifies three situations or purposes for reading that are covered in the NAEP reading assessment:

1) Reading for literary experience;
2) Reading to be informed; and
3) Reading to perform a task.

The stances define a second, cross-cutting dimension intended to capture the various ways in which readers respond to a given text as they read for meaning. The four stances articulated in the current framework are

1) Forming an initial understanding;
2) Developing an interpretation;
3) Personal reflection and response; and
4) Demonstrating a critical stance.

In scoring the NAEP reading assessment, three subscales based on the three reading situations are used. That is, each item uniquely contributes to the achievement estimate for a particular situation-based subscale and results are reported separately for each of these subscales, as well as for an overall composite scale.[4] Subscale reporting allows one to determine, for example, whether students in grade four are as adept in reading for information as in reading literary works.

Separate subscales are not reported for the four stances. However, each item is classified according to the stance it is intended to measure, and some analysis of differential student performance across stances can be accomplished using item-level results.

[3] NAEP Reading Consensus Project, *Reading Framework for the 1992 and 1994 National Assessment of Educational Progress* (Washington, D.C.: National Assessment Governing Board, 1993), 11-12.

[4] As explained later in the chapter, there are only two subscales at grade four: reading experience and reading for information.

## Studies Conducted by the Panel

As noted in the introduction to this chapter, the reading framework, as well as the 1992 item pool, was evaluated by the Panel in conjunction with the 1992 TSA. For the evaluation of the framework the Panel sought the advice of a number of reading experts, including a diverse and respected group of 21 experts who were interviewed shortly after the release of the 1992 reading results. The latter were asked to comment on the strengths and weaknesses of the framework as a basis for a national assessment, to consider whether the framework adequately represented current theory and practice in the field of reading, and to estimate the length of time before the framework would have to be revised to reflect changing conceptions of reading and reading assessment.[5]

In 1994, the Panel's evaluation of the assessment's content validity focused primarily on the 1994 item pool. To inform the evaluation, the Panel convened a group of reading experts, listed in appendix B, who conducted the analyses of the item pool commissioned by the Panel. In the first of these analyses, the reading experts classified each of the items into the situation and stance categories that, in their opinion, best characterized the item. They then evaluated the extent to which their classifications matched those assigned by the NAEP contractor. The outcomes of this analysis provided evidence of the clarity and utility of the framework dimensions as well as of the success of the item development process. In a second analysis, the Panel considered how well the distribution of tasks, determined by the expert advisors' classifications, matched the distributions recommended by the framework. Third, the Panel considered the difficulty distribution of the item pool and the implications of this difficulty distribution for appropriate measurement of reading achievement in the nation and the states. Finally, the Panel, with the assistance of its expert advisors, considered more generally the extent to which the 1994 item pool comprised reading tasks that were meaningful, appropriate, and valid measures of reading achievement.[6]

## The Panel's Findings

*In general, the 1994 studies, like those conducted for 1992, confirmed the validity of the NAEP reading assessment and the framework on which it was based, finding them reasonably well aligned with current research as well as with common classroom practices.* The Panel's expert advisors approved of the passages selected for the students to read and considered the questions that followed them to be appropriate. In particular, the experts concluded that, within the present limitations of on-demand, large-scale assessment design, the passages and items were generally successful in

[5] J.H. Mitchell, "Evaluation of the 1992 Reading Framework for the National Assessment of Educational Progress," in *The Trial State Assessment: Prospects and Realities: Background Studies* (Stanford, CA: The National Academy of Education, 1993).

[6] L. DeStefano, D. Pearson, and P. Afflerbach, "Content Validation of the 1994 NAEP in Reading: Assessing the Relationship Between the 1994 Assessment and the Reading Framework," in *Quality and Utility: The 1994 Trial State Assessment in Reading, Background Studies* (Stanford, CA: The National Academy of Education, forthcoming).

modeling the kinds of real-world reading tasks that students should be able to perform. The experts also approved of the emphasis on constructed-response items (about 70 percent of testing time was devoted to such items), believing that these items capture important aspects of the reading process that cannot otherwise be measured.

With regard to the reading framework itself, the expert advisors reaffirmed that the framework's general model of reading for meaning was consistent with current research and practice and worked well as the basis for an assessment. They also approved the developers' efforts to articulate the multidimensionality of the reading process and to report achievement separately for the different types of reading defined by the three reading situations.

Despite this general approval of both the framework and the tasks, the Panel, after reviewing the observations of the reading experts and other data, noted a number of areas in which the assessment could be improved. With respect to the *pool of 1994 assessment tasks,* the problems noted include

- Omission of items measuring reading to perform a task at grade four;
- Deficiencies in the scoring guides used to assign credit for constructed-response items;
- An overall preponderance of items in the item pool that are difficult for the majority of students who take the assessment; and
- Too few items that adequately capture the essential features of advanced reading achievement.

With regard to the *framework,* the Panel and its experts acknowledged that modeling the dimensions of the reading process for assessment purposes is a difficult problem requiring more research and suggested two areas that should be given particular scrutiny when the framework is next revised. These are

- Clarification and possible revision of the stance dimension and
- Expansion of the organizing dimensions to encompass additional components of the reading process.

The following paragraphs provide more detail regarding these concerns.

### *Omission of Items Measuring Reading to Perform a Task at Grade Four*

The developers of the reading framework wanted to define an assessment in which the relative emphasis given to each of the reading situations in the overall reading score was an appropriate reflection of the kinds of reading tasks undertaken by students at the targeted grade levels. They addressed this goal by specifying the proportion of the item pool that should be devoted to each situation at each grade level. Though this method seemed to work reasonably well at the 8th and 12th grades, operationalizing it for the 4th grade proved more problematic.

The difficulty stemmed from an apparent mismatch between the number of items necessary to construct a reliable subscale and the estimated proportion of time that fourth graders spend reading to perform a task. Because of the overall size of the item pool, the framework developers estimated that at least 20 percent of the items in the pool would have to be classified as representing a given situation type in order for the resulting subscale to be reliably determined. They also concluded, however, that "reading to perform a task, though important, probably does not reflect one-fifth of the type of independent reading fourth graders usually do." In order to avoid either over-representing this situation for fourth-grade readers or creating an unreliable subscale, the developers decided to omit this type of item entirely at grade four.[7] This decision resulted in the distribution of reading tasks by situation shown in table 2.1.

**Table 2.1** ***Framework for the 1994 reading assessment: percentage distribution of tasks by grade and reading situation***

| Grade | Reading Situation | | |
|---|---|---|---|
| | **Literary experience** | **To be informed** | **To perform a task** |
| 4 | 55% | 45% | (No Scale) |
| 8 | 40% | 40% | 20% |
| 12 | 35% | 45% | 20% |

SOURCE: NAEP Reading Consensus Project, *Reading Framework for the 1992 and 1994 National Assessment of Educational Progress* (Washington, D.C.: National Assessment Governing Board, 1993), 12.

*The reading experts convened by the NAE evaluation Panel reviewed this decision and concluded that dropping the subscale was not advisable. Rather, they held that the omission of items measuring reading to perform a task incorrectly implies that reading to perform a task is outside the scope of fourth-grade readers.*

***The Panel therefore suggests that when the reading framework is revised, content and measurement experts collaborate to consider alternative methods for maintaining both subscale reliability and age-appropriate distributions of reading tasks.*** It should be possible, for example, to take advantage of the fact that under current procedures the emphasis given to different types of reading in the summary measure of reading achievement is not entirely dependent on the task mix in the original item pool. Instead, as noted above, the contractor first scores the items onto the separate subscales. The overall reading score is then derived by giving each subscale a different weight or emphasis, appropriate to the grade level, and then combining the weighted subscores. Reliable measurement of each subscale thus could be obtained by representing the reading situations fairly equally in the item pool, while still maintaining an age-appropriate emphasis on types of reading through the subsequent weighting process. In this way, the need to drop a subscale at any given grade level could be avoided.

[7] NAEP Reading Consensus Project, op. cit., 12.

When students write their own answers to assessment questions, human scorers must evaluate the adequacy of each unique response. For large-scale assessments, written scoring guides are developed to assist the scorers in this process. The guides help ensure score reliability by specifying the kinds of responses that are acceptable for each question.

NAEP scoring guides undergo a multistage process of development and revision. First, NAEP developers create preliminary scoring guides at the time they write each item. These guides are subsequently tested against actual student responses and refined. One aspect of the refinement process is the incorporation of selected student responses into the guides as exemplars. These exemplars help convey the intent of the various score levels to those who will perform the operational scoring. Finally, during the scorers' training sessions, the guides are further elaborated to address significant questions or problems that arise in the interpretation of student responses. Appendix A contains the scoring guides for the two constructed-response items that were presented in figure 2.1. Notice that the guide for the second question, which is intended to elicit a longer and more detailed response from the student, identifies four different score levels, whereas the guide for the first question, which is intended to be answered more succinctly, contains only two levels.

*In reviewing the 1994 item pool, the Panel's expert advisors concluded that the quality of the scoring guides was uneven and that there were too many instances in which the guides were internally inconsistent or poorly aligned with other features of the items.* Guides that evaluated responses on the basis of criteria that had not been specified in the directions to the student were particularly problematic. Thus, for instance, guides for some questions gave higher scores to students who supported their arguments with more examples from the text (e.g., three examples rather than one or two) despite the fact that the questions themselves had not directed students to construct their answers on that basis. A case in point is the scoring guide for the question that asks students to relate Spider or Turtle to someone they know, have read about, or have seen in the movies or on television. (See appendix A.) Here the distinction between Level 3 and Level 4 responses hinges largely on the *number of comparisons made,* despite the fact that this criterion is never specified in the wording of the question or in the directions for the assessment.

The scoring guides for items classified as "demonstrating a critical stance" were particularly problematic. According to the Panel's expert advisors, these guides frequently failed to reflect the more complex reasoning skills that one would expect from successful responses to items of this type. Instead, the scoring guides for some of these items seemed to concentrate on more quantifiable—but less essential—criteria, such as examples from the text. Finally, the Panel's experts identified guides that contained unacceptable inconsistencies within and between score levels. For example, the scoring guide for one item classified by NAEP as "developing an interpretation" assigned the same score for a response that listed two facts (which the Panel's experts considered to be evidence of "initial understanding") and for a

response that provided a thorough discussion of one important characteristic (a more appropriate instance of "developing an interpretation").[8,9]

Since conducting this analysis, the Panel has learned that, for new items that are added to the reading assessment, the NAEP contractor has already initiated efforts to improve the scoring guides and the clarity of directions to students. The Panel recommends that these efforts be continued both with respect to items written for immediately upcoming assessments and in preparation for assessments that will follow the next framework revision.

## *Problems with the Distribution of Item Difficulties*

According to the framework, the assessment should be based on reading passages, or texts, that "range in difficulty from those that specific grade-level teachers agree could be read by the least proficient students in a class (e.g., about second-grade level in a fourth-grade class) to those texts that can be read by only the most proficient readers in the class (possibly eighth-grade level in a fourth-grade class)."[10] Furthermore, the framework states that the assessment should consist of items or tasks "that most students at the given grade levels can do. This means not only that students possess the requisite abilities, but also that they are likely to have actually encountered the particular type of text or task."[11]

*It appears, however, that with regard to the fourth grade, the actual difficulty of the reading assessment tasks is shifted upward relative to the current achievement of students. The assessment contains a large number of tasks that relatively few students can answer correctly and relatively few tasks that are within the scope of the less able students.*[12] Because student achievement estimates are developed through a scaling process rather than by a simple addition of the number of correct items, this does not necessarily mean that it is harder to get a "good score" on the assessment.[13] It does

[8] The framework defines the four stances as follows: 1) initial understanding requires the reader to provide an initial impression or unreflected understanding of what was read; 2) developing interpretation requires the reader to go beyond the initial impression to develop a more complete understanding of what was read; 3) personal reflection and response requires the reader to connect knowledge from the text with his or her own personal background knowledge; and 4) demonstrating a critical stance requires the reader to stand apart from the text and consider it. (NAEP Reading Consensus Project, op. cit., 16-17).

[9] Similar problems with scoring guides were identified by a different set of expert advisors who evaluated the 1992 reading item pool on behalf of the Panel. See D. Pearson and L. DeStefano, "Content Validation of the 1992 NAEP in Reading: Classifying Items According to the Reading Framework," in *The Trial State Assessment: Prospects and Realities: Background Studies* (Stanford, CA: The National Academy of Education, 1993). This provides some validation for the judgment of the 1994 panelists because there was considerable overlap between the items in the 1992 and 1994 assessments. In fact, fewer scoring guide problems were identified among the items that appeared for the first time in 1994.

[10] NAEP Reading Consensus Project, op. cit., 20.

[11] Ibid., 20.

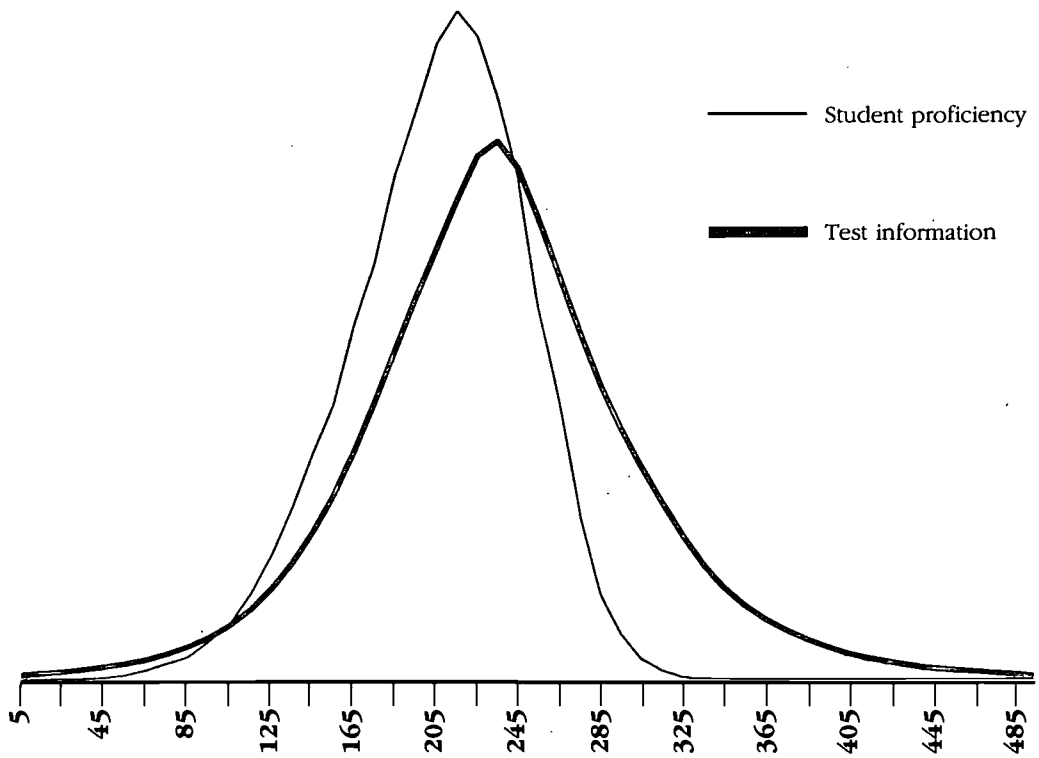
[12] The situation is much less acute at grades 8 and 12 because of the practice of using a proportion of the item blocks across adjacent grades. Thus the 8th-grade assessment includes two item sets that are also suitable for 4th graders and the 12th-grade assessment includes two sets that are also used at grade 8.

[13] This language is somewhat misleading because no individual student actually gets a "score" on the NAEP assessment.

mean, however, that it is more difficult to get precise and reliable estimates of achievement for less able students. Additionally, some students may find the assessment excessively frustrating, even to the point that they stop trying to do well on it.

Figure 2.2 illustrates this problem of item difficulty shift for the case of the 1994 fourth-grade reading for information subscale. The lighter curve shows the distribution of student proficiency estimates. Note that the greatest number of students have proficiency estimates around 205 on this subscale, whereas virtually none have estimates below 85 or above 325. The difficulty distribution of the item pool, however, is such that the greatest accuracy of estimation occurs around 265; this is shown by the darker, test information curve.[14] The curve further shows that the

**Figure 2.2.** ***Comparison of the distributions of student proficiency and test information in relation to the NAEP achievement scale: 1994 fourth-grade reading assessment, reading for information subscale***




SOURCE: Unpublished Educational Testing Service data

NOTE: For any given point on the achievement scale, the height of the student proficiency curve indicates the proportion of students whose reading achievement is estimated to be at that point, and the height of the test information curve indicates the degree of accuracy with which achievement can be estimated.

[14] Using Item Response Theory Scaling, each NAEP item is associated with a particular range on the achievement scale, related to the difficulty of the item. Accurate estimation of a student's proficiency requires that the student be tested on several items grouped in a difficulty range that is challenging, but not impossible for that student. These closely-spaced multiple measures help to pinpoint the student's true proficiency. When the student responds to items that are mostly too easy or mostly too hard, there is less information from which to judge his or her true proficiency.

amount of information provided by the test is sufficient even at the very top of the effective proficiency distribution for fourth-grade students (around 325), but the amount of information tails off substantially for students in the lowest part of the distribution (below 100 on the achievement scale). A similar pattern occurs on the second reading for literary experience subscale (not shown).

It is important to recognize that this upward shift in difficulty of NAEP assessments relative to the population has occurred not only in reading. In fact, it has been much more pronounced in some of the other subject areas for which NAEP has developed new content frameworks since 1990. No careful analysis has yet been conducted to ascertain the specific causes of the phenomenon. However, one explanation might be that the frameworks provide only very loose guidelines to help item writers predict the difficulty of texts and tasks. Once the *actual* difficulty is known, based on the field test results, there is only limited opportunity to modify the difficulty distribution of the pool; most of the statistically sound items must be retained regardless of difficulty. An alternative explanation is that most American students may simply lack experience with the kinds of complex tasks demanded by our national education goals.

The Panel recommends that the NAEP contractor adjust the distribution of easy and hard items so that there are more items that can be successfully attempted by the majority of American students, as well as more items with difficulty levels appropriate for lower-achieving students. It is important to note that this does not mean returning to the assessment of isolated skills. Rather, the easier items, like the more difficult ones, should still allow students to demonstrate their proficiency in each of the reading situations and on each of the stances identified by the framework.

Because the achievement levels established by NAGB are set at levels that are challenging for the average student, expanding the numbers of easier items will not improve the precision of estimates for the percentages of students achieving advanced, proficient, or even basic levels of proficiency.[15] However, describing the performance and progress of students with lesser proficiency is also an important goal for the nation's key indicator of educational progress and provides significant information for policy makers, educators, and the public. Furthermore, motivation is increased and equity upheld if every student who is asked to participate in NAEP is offered a least some items at difficulty levels that challenge, but do not frustrate.

### *Too Few Items that Capture the Essential Features of Advanced Reading Achievement*

In order to appropriately measure advanced performance, it is not sufficient for assessment tasks to be simply "hard" in the statistical sense. Rather, these hard tasks must also evidence mastery of the particular reading outcomes that are considered to be most critical to successful reading achievement. The NAGB achievement levels, for example, describe advanced performance for fourth-grade students as the ability "to generalize about topics in the reading selection and demonstrate an awareness of how authors compose and use literary devices. When reading text appropriate to fourth

[15] In 1994, an estimated 40 percent of fourth-grade students had estimated reading proficiency scores that fell below the cutoff for the basic achievement level.

grade, [these students] should be able to judge texts critically and, in general, give thorough answers that indicate careful thought."[16] *The deliberations of the Panel's expert advisors indicate, however, that there are relatively few items in the 1994 reading item pool that successfully tap these challenging outcomes or exemplify the requirements of advanced achievement.*

At one level, the shortage of advanced items is related to fundamental constraints of the NAEP design. In reading, as in a number of subject areas, many of the competencies that distinguish the most able students may only be evident in sustained performances that require far more time than the few minutes per item allowed in the present NAEP assessment format. Yet designing longer assessment tasks creates its own problems because reliable scores cannot be obtained unless each student interacts with a reasonably large number of assessment tasks. This dilemma reflects a major unsolved research problem that affects most large-scale assessments including NAEP, college and professional school entrance examinations, and the majority of state assessments.

At another level however, the Panel's experts agreed that there were at least some NAEP reading items that did a reasonable job of addressing advanced performance. ***The Panel suggests that the NAEP contractor, in collaboration with its own expert advisors, undertake additional research to better understand the essential features of the most exemplary advanced items (considered also in the context of how these items actually performed when presented to NAEP examinees) and to generate additional advanced items to the same model.***

Finally, the Panel considers that the assessment of both the most and least proficient students would be facilitated if students could be better matched to assessment tasks. Under the current NAEP design, each student sees a more or less randomly representative subset of the items. This makes it difficult to gather much useful information from student whose reading performance is near the very bottom or the very top of the distribution because nearly all of the items in their assessment books will be either too hard or too easy for them. The affected students include many of the students with disabilities or limited English proficiency who are being increasingly included in the assessment,[17] as well as those who are the most advanced performers. Conversely, items that are written especially for either of these groups of students may have limited utility for measuring the achievement of many of the other students in the NAEP sample, particularly those whose performance is nearer the opposite end of the distribution.

The Panel therefore recommends that NAGB and NCES explore the feasibility of adaptive testing methodologies that would allow students to be matched with item blocks appropriate to their level of competence. This recommendation will be developed further in the Panel's forthcoming capstone report.

[16] J.R. Campbell, P.L. Donahue, C.M. Reese, and G.W. Phillips, *NAEP 1994 Reading Report Card for the Nation and the States* (Washington, D.C.: National Center for Education Statistics, January 1996), 42.

[17] See chapter 4 of this report.

## *Lack of Clarity in the Stance Dimension*

*When the Panel's expert advisors independently classified the items in the 1994 item pool according to the two dimensions specified in the framework, their classifications successfully matched the item to the* ***situation*** *coded by the NAEP contractor (reading for literary experience, reading for information, or reading to perform a task) in every case.* This is an important, positive finding because, as discussed above, the situation dimension defines reporting subscales. The high level of agreement suggests that there is a common understanding of what is being reported on each of these subscales.

*With regard to the* ***stance*** *dimension however, agreement between the expert advisors and NAEP was much lower: across all four stances and all three grades, average agreement was only 67 percent.* This does not affect subscale reporting because there are no subscales built around the stances, but it does suggest that the stance dimension adds little interpretive value to the assessment results.

Table 2.2 displays the specific pattern of agreements and disagreements about the stance classifications between the Panel's expert advisors and the NAEP contractor. The percentages along the unshaded diagonal indicating agreement. The highest rate of agreement, 79 percent, pertains to items classified by NAEP as "developing an interpretation"; the lowest rate, 46 percent, is for items classified as "demonstrating a critical stance." In the latter case, a high proportion of the items that NAEP designated as critical stance were classified by the Panel's experts as developing interpretation. Many of these disagreements hinged on the scoring guides that, as discussed above, too often focused on quantifiable features of the student responses rather than the more essential characteristics of critical stance.

More generally, the Panel's experts felt that the high rate of classification disagreements might be an indication that the reading stances specified by the framework are not cognitively discrete. Responding to an assessment task sometimes involves two or more stances, and the overlapping of stances that naturally occurs while reading for meaning adds complexity to the classification process. As one of the expert advisors explained, "in effect the framework tries to 'freeze frame' the reading process at a particular point to determine things like initial understanding, but no reader actually (intentionally) would stop reading at that point...there is an artifice of trying to deconstruct the act of reading to accommodate the framework."[18]

Additional ambiguity may have been introduced into the classification process by the lack of more precise stance definitions in the reading framework. Currently, the framework defines the four stances almost entirely by example: that is, the framework lists the kinds of questions one would ask of students in order to elicit a particular stance, but does not otherwise describe the stances. Furthermore, some of the example questions given in the framework did not seem particularly clear to the Panel's expert advisors. For instance, some felt that the question "What does this passage/story say to you?" was more indicative of an initial understanding stance than of the personal response category it was intended to illustrate.[19]

[18] L. DeStefano, D. Pearson, and P. Afflerbach, op. cit., 19.

[19] NAEP Reading Consensus Project, op. cit., 14.

**Table 2.2.** ***Classification of items into stances: expert advisor classifications compared to official NAEP classifications (Grades 4, 8 and 12 combined)***

| Expert advisor classifications | Official NAEP classifications | | | | TOTAL |
|---|---|---|---|---|---|
| | Initial Understanding<br>N<br>% of column total | Developing Interpretation<br>N<br>% of column total | Personal Response<br>N<br>% of column total | Critical Stance<br>N<br>% of column total | N<br>% of column total |
| Initial Understanding | 20<br>71% | 24<br>16% | 1<br>2% | 4<br>4% | 49<br>14% |
| Developing Interpretation | 7<br>25% | 120<br>79% | 9<br>16% | 48<br>46% | 184<br>54% |
| Personal Response | | 2<br>1% | 39<br>71% | 4<br>4% | 45<br>13% |
| Critical Stance | 1<br>4% | 5<br>3% | 6<br>11% | 48<br>46% | 60<br>18% |
| TOTAL | 28 | 151 | 55 | 104 | 338 |

SOURCE: L. DeStefano, D. Pearson, and P. Afflerbach, "Content Validation of the 1994 NAEP in Reading: Assessing the Relationship Between the 1994 Assessment and the Reading Framework," in *Quality and Utility: The 1994 Trial State Assessment in Reading, Background Studies* (Stanford, CA: The National Academy of Education, forthcoming).

## *Unaddressed Components of the Reading Process*

Finally, the Panel's expert advisors noted that there are important aspects of the reading model presented in the framework that are not captured in the organizing structure of situations and stances. In a statement with which most reading experts would agree, the framework describes reading for meaning as "a dynamic, complex interaction among three elements: the reader, the text, and the context...good readers bring to this interaction their prior knowledge about the topic of the text and their purposes for reading it."[20] However, while the situation dimension accounts for type of text and certain aspects of reading purpose, there are no separate dimensions to deal with variations in the student's prior knowledge or the contextual factors associated with cultural differences.

To illustrate, consider, for example, that just as one uses somewhat different strategies and skills when reading for information versus reading for literary experience, so does one address a reading task differently if one is exploring a new topic or discipline

[20] Ibid., 10.

versus reading for incremental knowledge or for a new perspective in a familiar domain. In theory, it would be possible to add a dimension associated with different levels of prior knowledge and then to design separate reading tasks that measure reading proficiency under different conditions of prior knowledge. One would need, however, a method for pairing tasks appropriately with examinees who are or are not familiar with the topic of the reading text.

Improving the measurement of reading competence for students of diverse culture, background, or experience is another important goal that has proven difficult to solve in large-scale assessments wherein all examinees traditionally respond to a uniform set of tasks. NAEP, like most similar assessments, incorporates several strategies to minimize the impact of individual diversity on scores. First, assessment tasks are restricted to those that require minimum prior knowledge (except insofar as such knowledge is essential to the subject under assessment). Second, these tasks are based on a wide range of stimulus materials (e.g., many different kinds of reading passages) so that each examinee will have a greater chance of finding at least some tasks that cover familiar topics or present familiar perspectives. Third, all tasks are screened during development to eliminate those that, when presented to examinees of generally equivalent proficiency, are particularly easy or hard for students of one gender or from one particular background. These measures do help prevent the more egregious instances of bias; they do little, however, to actually incorporate into the assessment the interaction of reading proficiency with the contextual factors of background, experience, or interests.

## *Summary and Recommendations*

*In summary, the Panel, upon reviewing the evidence from its commissioned reading experts and others, concluded that the 1994 assessment provides an appropriate measure of reading achievement and one that is reasonably congruent with current reading theory and common classroom practices. The assessment includes a number of desirable features as well as some for which further improvements are clearly needed, either now or when the framework is next revised. None of these problems, however, is so troublesome as to have undermined the validity of the 1994 assessment.*

Based on these findings, the Panel recommends that the general structure (framework) of the present reading assessment be maintained through the year 2000 or 2002, while staged efforts are meanwhile undertaken to enhance the quality of the assessment and the value of the results. This general recommendation is in keeping with both the *relative stability principle,* which encourages that subject area frameworks be maintained long enough to sustain meaningful trend measurement, and the *quality principle,* which calls upon NAEP to be exemplary in the development and use of assessment techniques.

Below, the Panel reviews briefly the recommendations and suggestions stemming from its preceding analyses. The chapter ends with the Panel's broader observations and recommendations concerning the type of framework and assessment development process that will be required to successfully implement the suggested improvements in reading.

The problem areas identified in the current reading assessment can be divided into two main groups: those that can be remediated immediately without compromising trend measurement, and those that require more time—either because implementation before the institution of a new framework would jeopardize the trend line, or because further research and development is needed to identify solutions.

Recommendations for immediate implementation:

The Panel recommends that the National Center for Education Statistics (NCES) and its contractor take the following actions in preparation for the next NAEP reading assessment:

- Improve the difficulty distribution of the item bank by including more items that can be successfully attempted by the majority of students as well as easier passages and items appropriate to lower-achieving students. The Panel restates that this does not mean returning to the assessment of isolated skills. Rather, items should be constructed so that lower-achieving students may still demonstrate their proficiency in each of the reading situations and on each of the stances identified in the framework.

- Continue work to enhance the consistency of the scoring guides and student directions.

- Develop and include more items that adequately measure the essential features of advanced reading achievement.

Suggested areas for further analysis and research:

Furthermore, the Panel recommends that work begin now to lay the foundation for an even stronger framework and assessment in anticipation of the time when the reading framework is next revised. Specific areas for analysis and research deriving from the preceding discussions include

- ***Developing better guidelines for predicting passage and item difficulty so that the difficulty distribution of the item pool can be better aligned with actual student proficiency.***

- ***Investigating alternative methods for achieving the concurrent goals of subscale reliability and age-appropriate distributions of reading tasks.***

- ***Refining the organizing structure in the framework so that the dimensions it describes are clear, cognitively distinct, and faithful to the way readers actually read. These refinements***

***would include reviewing and revising the stance dimension and incorporating new dimensions to take account of differences in prior knowledge and of the contextual factors associated with differences in background, experience, and interests.***

- ***Investigating additional methods for assessing advanced achievement that expand the current limitations of large-scale assessment while preserving reliability and cost-efficiency.***

## *Implementing the Panel's Recommendations: Toward a More Effective and Efficient Assessment Development Process*

The Panel would be remiss in advancing the above recommendations without also acknowledging the constraints under which NAEP developers currently work. We therefore conclude this chapter with some observations about those constraints and recommendations for improving the development process. These issues will be further explored and expanded in the Panel's summative capstone report.

*Currently, with regard to reading, as with other subject areas, structural features of the NAEP funding and development process leave little time for the kind of planned, farsighted content development on which NAEP, as the key indicator of what the nation's students know and can do, ought to be based. Hemmed in by short authorization and funding cycles on the one hand, and time consuming federal clearance procedures on the other, the actual development of frameworks and assessment tasks has been squeezed into unconscionably short time frames.* In particular, during years when new frameworks are adopted, the NAEP contractor has typically had less than six months in which to develop field test materials before these materials must be finalized and sent to the Office of Management and Budget for clearance. During this period, the contractor must develop two or three times the number of items needed for an actual administration in order to allow for items that will be rejected during the expert review process or that will fail statistically in field test. The implications of such an overly rushed development cycle can be better appreciated when one realizes that assessment tasks that are produced during this one brief period will set the tone for all future assessments until the next revision of the framework, eight or ten years in the future. Trend measurement (in addition to cost considerations) dictates that high proportions of items be retained from cycle-to-cycle, and even those items that are replaced between cycles must remain consistent with the earlier items. Thus, for example, only about 20 percent of the tasks in the fourth-grade reading item pool were replaced between 1992 and 1994, and the new materials are only modestly different from those they replaced.

Whereas a high volume of items can be generated in a short time frame by hiring sufficient editorial staff and item writers, only the addition of more time will enable the kind of iterative development process that can ensure high quality assessment content. Because NAEP item developers must grapple with the challenging constructs and innovative item formats required in the new frameworks, it is especially important for development to progress through a series of small-scale pilot tests, reviews, and revisions—a process antithetical to the mass production and one-chance field tests

currently employed. Ideally, framework development should also be integrated into this iterative process, thereby allowing new ideas to be tried out empirically before being codified into the final framework. Had such procedures been in place when the current reading assessment was under initial development in 1990 and 1991, it is far less likely that the Panel's reading experts would have voiced as many concerns about the lack of clarity of the stance dimension, the uneven quality of the scoring guides, or the shortage of items measuring the essential features of advanced reading outcomes.

The Panel recommends that, for every NAEP subject area, NAGB and NCES adopt a process that allows new research and development to begin several years before the framework is scheduled to be revised and a new trend line begun. This research and development could progress in a relatively modest manner through successive pilot studies and small-scale trials targeted at particularly challenging research problems. When it is time to begin the actual revision, Congress, NAGB, and NCES should allow for a framework and item development cycle that is substantially longer and more integrated than the current one. The Panel further recommends that Congress forward fund NAEP in order to facilitate this process.

# 3 Sampling and Assessment Administration for the 1994 TSA

## Introduction

In 1994, as in its previous evaluation cycles, the Panel again considered the sampling and administration procedures used in conducting the TSA. The high quality of TSA sampling and administration is essential to ensuring representative and comparable information across the states and, ultimately, fair and accurate results. In addition, the success of school and student recruitment and the smoothness of the administration speak to the feasibility of continuing state assessments. Consequently, considerations of these factors remains a critical component of the Panel's evaluation.

In *Assessing Student Achievement in the States,* the Panel's first report to Congress, the Panel concluded that the conduct of the 1990 TSA was generally successful in every category—the sampling design met the high standards expected of a major national education survey, the participation rates of schools and students selected for the sample were generally good, and the assessment appeared to have been carefully and consistently administered for nearly all students who participated.[1]

The 1990 trial had been deliberately limited in size because of the many uncertainties associated with this new undertaking. Consequently, only eighth-grade mathematics was assessed. In 1992, the size of the assessment was increased by a factor of three, with mathematics assessed at both grades four and eight and reading added at grade four. As a result, the 1992 trial offered an opportunity to evaluate the robustness of participation rates and assessment procedures under the burden of a larger assessment.

The Panel's conclusions regarding the 1992 TSA generally reflected the positive findings from 1990. However, the extent of the burden on states and schools became more apparent when, with the expansion to two subjects at grade four, the fourth-grade sample in some smaller jurisdictions began to approximate the total number of available schools and, in some cases, the total number of students. The Panel also noted that participation rates for sampled schools, which had been extremely high in 1990, declined in 1992 except in those states that had mandated participation by designating the TSA as part of their own state assessments. Finally, a statistically significant, negative relationship between school participation rates and mean state achievement scores was observed, suggesting the possibility that higher average performance in some states might be related to the differential participation of schools with more able students.[2]

The original plans for 1994 called for a further expansion of the TSA, possibly to match the three subjects/three grades design of national NAEP assessments.

[1] The National Academy of Education, *Assessing Student Achievement in the States* (Stanford, CA: Author, 1992), 7-8.

[2] The National Academy of Education, *The Trial State Assessment: Prospects and Realities* (Stanford, CA: Author, 1993), 45-47.

However, funding constraints and competing priorities from other components of the NAEP program forced the 1994 TSA to scale back to a single subject and grade—fourth-grade reading. Table 3.1 summarizes the grades, subjects, and the numbers of participating jurisdictions, schools, and students for each of the three TSAs.

**Table 3.1.** ***Comparison of 1990, 1992, and 1994 Trial State Assessments***

| | 1990 TSA | 1992 TSA | 1994 TSA |
|---|---|---|---|
| Grades assessed | 8 | 4 and 8 | 4 |
| Subject areas assessed | Mathematics | Mathematics (Grade-4 and -8)<br>Reading (Grade 4 only) | Reading |
| Number of participating states and other jurisdictions | 40 | 44 | 44 |
| Average number of participating schools per jurisdiction | 100 | 225 | 108 |
| Total number of participating schools | 3,600 | 8,700 | 4,732 |
| Total number of sessions | 3,900 | 13,900 | 4,871 |
| Total number of participating students | 101,000 | 331,000 | 121,694 |

SOURCES: National Center for Education Statistics, *1990 Trial State Assessment: Summary of Participation Rates, Training, and Data Collection Activities* (Washington, D.C.: Author, September 1990; National Center for Education Statistics, *1992 Trial State Assessment: Summary of Participation Rates, Training, and Data Collection Activities* (Washington, D.C.: Author, July 1992); National Center for Education Statistics, *1994 Trial State Assessment: United States Report on Data Collection* (Washington, D.C.: Author, October 1994).

The reduced size of the 1994 TSA unfortunately limited the Panel's ability to draw summary conclusions regarding the feasibility and impact of a full-scale state assessment (i.e., an assessment spanning three subjects at three grades, or even two subjects at two grades). The Panel considered it important, however, to replicate its previous analyses of public school sampling and participation, the conditions of administration, and the comparability of average performance between matched groups of students taking the state and national assessments. In addition, there were significant new features of the 1994 TSA that required special attention—particularly the first-time addition of nonpublic schools to the sample—and a number of specific research questions that had arisen out of the 1992 evaluation.

The following areas received special emphasis in the Panel's evaluation of the 1994 TSA:

- Sampling, participation, and administration in nonpublic schools;
- The impact of public school nonparticipation on mean state achievement scores;
- The nature and significance of irregularities in session administration procedures; and
- The exclusion and assessability of students with disabilities or limited English proficiency (this topic is treated separately in chapter 4).

The chapter is organized into three main sections. The first examines issues related to the sampling and participation of public schools and public school students. The second section addresses the same topics for nonpublic schools and their students. Finally, the third section presents information about the administration of the TSA, and includes discussion of the comparability of student performance between state and national NAEP, the comparability of student performance between sessions that were or were not monitored by representatives of the contractor, and the frequency and impact of session irregularities.

## *Public School Samples*

### *Sampling and Recruitment of Schools*

The sampling and recruitment of public schools in 1994 generally paralleled the experience of the earlier TSAs. The target population was made up of all fourth-grade students in regular public and nonpublic schools in the participating states and other jurisdictions, except for those IEP and LEP students who were judged incapable of taking the test.[3] For each state, the sampling design required approximately 100 public and 15 nonpublic schools from which approximately 2,500 students were then drawn. The state lists from which the school samples were drawn were based on a commercially available data base, Quality Education Data (QED), and checked against NCES records. Procedures were also implemented to identify new public schools not yet on the QED and add them to the list.

For public schools in all but the least populous jurisdictions, the contractor used a stratified sampling procedure: a systematic means for selecting schools that ensures that the resulting sample will reflect the overall composition of the state on certain key variables. Stratification was based on measures of median income and urbanization

[3] "Regular" schools are non-special education schools. Nonpublic schools are Catholic schools, other private schools, Bureau of Indian Affairs schools, and domestic Department of Defense schools. The overwhelming majority (97.4 percent) of students in the nonpublic category were drawn from Catholic and other private schools. For most jurisdictions, therefore, "nonpublic" is essentially identical with "private." The "other jurisdictions" included Washington D.C. (which withdrew after the data collection phase), Guam, and the Department of Defense Education Activity (DODEA) Overseas Schools.

for the communities in which the schools were located and on the percent minority enrollment of the schools. For the least populous jurisdictions with small numbers of schools (Delaware, the Department of Defense Overseas Schools, and Guam), all schools were included in the sample.

As in previous TSA cycles, recruitment of public schools was carried out by designated state department of education employees. These individuals, referred to as "state coordinators," worked from lists of originally sampled and substitute schools provided by Westat, the NAEP data collection contractor. State coordinators first recruited originally sampled schools and, if any declined, attempted to recruit substitute schools. Substitute schools were chosen to be as similar as possible to the refusing schools with regard to estimated grade enrollment, median household income, percent black enrollment and percent Hispanic enrollment. In cases in which no suitable substitute could be found among those schools not sampled (most often because all or most schools were included in the original sample), a school already in the sample was asked to conduct a second session, using a second sample of students.

Although there is no evidence that the procedure used for recruiting replacement schools actually had a detrimental effect on the 1994 sample, the Panel was troubled due to the fact that, unlike previous TSA procedure, 1994 state coordinators received their lists of originally sampled and substitute schools simultaneously.[4] This change was intended to enhance overall participation by allowing states to begin recruiting replacements while some originally sampled schools were still pending, particularly in cases where districts took a long time deliberating their participation decisions. It also, however, raised the potential for bias (i.e., the potential that the final sample would not be representative of the whole) because states, upon seeing their alternatives, might selectively recruit those schools, from each original/substitute pair, that appeared more likely to cooperate or more likely to perform well in the TSA. Allowing recruiters to know the identities of the substitute schools before the original schools had declined was unorthodox and opened the state assessments to criticism even if no improprieties actually occurred. Doing so was analogous to allowing assessment administrators to open the packages of assessment booklets in advance, a practice that is routinely forbidden even though most, if not all, administrators would likely maintain the security of the assessment tasks.

The Panel commissioned a sampling expert to examine the methods by which schools and students were sampled for the 1994 TSA.[5] *After reviewing his report, the Panel concluded, as it had in previous TSA cycles, that the 1994 sampling design and execution for public schools were appropriate and generally met the high standards expected of a major national education survey. The one exception was the procedure for recruiting substitute schools that, as noted, could have resulted in differential nonresponse patterns and thus increased bias.*

The Panel recommended, and NCES adopted, alternative recruitment procedures for substitute schools in the 1996 state assessments. Under the new procedures, the

[4] Westat, *Report on Data Collection Activities for the 1994 National Assessment of Educational Progress* (Washington, D.C.: Author, March 1995).

[5] B.D. Spencer, "School and Student Sampling in the 1994 TSA: An Evaluation," in *Quality and Utility: The 1994 Trial State Assessment in Reading, Background Studies* (Stanford, CA: The National Academy of Education, forthcoming).

names of substitute schools were stored in the schools-list data base, but the state coordinator did not have access to the name of a substitute until he or she had permanently coded the originally sampled school as refusing.

## *School Participation Rates*

In general, the representativeness of the final sample can only be ensured by recruiting high percentages of the originally sampled schools. Less than perfect participation rates are a serious threat to representativeness because of the likelihood that refusing schools, as a class, may differ from cooperating schools in systematic ways that are not fully captured by the matching procedures used to select substitutes. When reviewing participation rates therefore, the Panel focused on rates for originally sampled schools. For each of the TSAs, the overall before-substitution public school participation rates are shown in table 3.2. The before-substitution public school participation rates for recent national NAEP assessments are included for comparison.

The overall 1990 TSA participation rate was 94 percent, markedly higher than national NAEP public school participation rates for grades four and eight. The latter, since 1983-1984, have ranged from 84 percent to 90 percent. The rate attained in the 1990 TSA may have been due to the "novelty effect" of the first state assessment, however, because the overall public school participation rate fell to 88 percent in the 1992 TSA, after which it rose only slightly, to 89 percent, in the 1994 TSA. Despite the drop from 1990, the 1992 and 1994 TSA public school participation rates are quite good for a survey of this type and certainly compare favorably with the rates obtained in national NAEP.

In recognition of the bias that can result from nonparticipation, NCES procedures require that the reporting of survey results be contingent upon adequate weighted participation rates.[6] Specifically, for the TSA, these standards specify that results from states (or other participating jurisdictions) with weighted before-substitution school participation rates below 70 percent not be published in any NCES reports, and results from states that have 1) before-substitution rates between 70 and 85 percent *and* 2) after-substitution rates below 90 percent be annotated to indicate that the rates were low enough to raise concerns about representativeness.

By these NCES minimum standards, the 1990 TSA was exceptional: only two states had weighted before-substitution public school participation rates below 85 percent, and none fell below 70 percent. (See table 3.3.) In 1992, there was much more variability. Though many states maintained very high participation, 17 states fell

[6] Weighted school participation rates are calculated after each school in the original sample has been weighted to reflect the number of students represented by the school. Thus the nonparticipation of a school that represents many students would have a greater impact on the participation rate than the nonparticipation of a school that represents fewer students. Generally, weighted NAEP participation rates have been slightly higher than unweighted rates.

**Table 3.2.** ***Before-substitution public school participation rates (percent of originally sampled schools participating) in recent national and state NAEP assessments***

| Assessment | Grades 3-4 | Grades 7-8 |
|---|---|---|
| 1983-84 national NAEP | 89 | 90 |
| 1985-86 national NAEP winter sample | 87 | 84 |
| 1985-86 national NAEP spring sample | 89 | 89 |
| 1990 national NAEP | 88 | 87 |
| 1990 TSA | n/a | 94 |
| 1992 national NAEP | 86 | 85 |
| 1992 TSA | 88 | 88 |
| 1994 national NAEP | 86 | 86 |
| 1994 TSA | 89 | — |

—Not available

SOURCES: A.E. Beaton et al., *Implementing the New Design: The NAEP 1983-84 Technical Report* (Washington, D.C.: National Center for Education Statistics, March 1987), 87; A.E. Beaton et al., *Expanding the New Design: The NAEP 1985-86 Technical Report* (Washington, D.C.: National Center for Education Statistics, November 1988), 71; E.G. Johnson & N.L. Allen, *The NAEP 1990 Technical Report* (Washington, D.C.: National Center for Education Statistics, February 1992), 69; S.L. Koffler et al., *The Technical Report of NAEP's 1990 Trial State Assessment Program* (Washington, D.C.: National Center for Education Statistics, April 1991), 55; E.G. Johnson & J.E. Carlson, *The NAEP 1992 Technical Report* (Washington, D.C.: National Center for Education Statistics, July 1994), 84; E.G. Johnson et al., *Technical Report of the NAEP 1992 Trial State Assessment Program in Mathematics* (Washington, D.C.: National Center for Education Statistics, April 1993), 97; E.G. Johnson et al., *Technical Report of the NAEP 1992 Trial State Assessment Program in Reading* (Washington, D.C.: National Center for Education Statistics, February 1994), 76; N.L. Allen, J.E. Carlson, and D.L. Kline, *The NAEP 1994 Technical Report* (Washington, D.C.: National Center for Education Statistics, forthcoming), 5-9; J. Mazzeo et al., *Technical Report of the NAEP 1994 Trial State Assessment Program in Reading* (Washington, D.C.: National Center for Education Statistics, December 1995), 80.

NOTE: School participation rates are unweighted.

below 85 percent and four of these fell below 70 percent on one or more of the subject-by-grade samples.[7] These numbers do not include two states that withdrew during the recruitment period because too many schools declined to participate.

In 1994, somewhat fewer states (12) fell below the 85 percent weighted before-substitution participation rate criterion for public schools. (Rates for nonpublic schools were calculated separately and are discussed later in the chapter.) The states

[7] Despite concerns about the representativeness of results for these latter states, NCES chose to report their data because the minimum standards for reporting had not been in place at the time that the 1992 participation agreements were signed by the states.

with low public school participation included two, Idaho and Michigan, that had rates below 70 percent and thus were not reported. As in 1992, there were also two other states, Illinois and Ohio, that withdrew during the recruitment process when they realized that they would not meet the participation criterion. Furthermore, in response to a survey of state assessment directors carried out on behalf of the Panel, only 24 percent of the 1994 respondents cited difficulties in recruiting public schools.[8] This figure, which is substantially down from the 48 percent of respondents who had reported such problems after the 1992 TSA, probably reflects the smaller scope of the recruitment effort when only one subject and one grade were being assessed.

**Table 3.3.** ***States with weighted before-substitution public school participation rates below 85 percent: 1990, 1992, and 1994 TSAs***

| TSA | States with weighted before-substitution participation rates below 85 percent | |
|---|---|---|
| | Percent of states | States (participation rates shown in parentheses) |
| 1990 - Grade-8 math | 5% | IL (78), OK (78) |
| 1992 - Grade-8 math | 37% | ME (62), AL (66), NJ (69), NE (75), NM (77), OH (77), MI (78), ND (78), IN (79), NH (80), MN (81), NY (81), PA (81), OK (82), MA (83) |
| 1992 - Grade-4 math | 37% | ME (57), NH (69), ND (73), AL (75), NM (75), IN (76), NJ (76), NY (78), OH (79), NE (80), MN (82), MI (83), RI (83), ID (84), PA (84) |
| 1992 - Grade-4 reading | 34% | ME (58), NH (68), ND (70), AL (76), NE (76), NJ (76), NM (76), IN (77), NY (78), OH (78), MN (81), ID (82), MI (83), RI (83) |
| 1994 - Grade-4 reading | 29% | MI (63), ID (69), NE (71), NH (71), TN (72), NY (75), WI (79), CA (80), ND (80), PA (80), RI (80), IN (83) |

SOURCES: I.V.S. Mullis et al., *The State of Mathematics Achievement* (Washington, D.C.: National Center for Education Statistics, June 1991), 437; I.V.S. Mullis et al., *NAEP 1992 Mathematics Report Card for the Nation and the States* (Washington, D.C.: National Center for Education Statistics, April 1993), 316-318; I.V.S. Mullis et al., *NAEP 1992 Reading Report Card for the Nation and the States* (Washington, D.C.: National Center for Education Statistics, September 1993), 247; J.R. Campbell et al., *NAEP 1994 Reading Report Card for the Nation and the States* (Washington, D.C.: National Center for Education Statistics, January 1996), 105.

NOTE: School participation rates are weighted: they are calculated after each school in the original sample has been weighted to reflect the number of students represented by that school. Non-state jurisdictions are not included in this table, but none had rates below 85 percent.

[8] E. Hartka, J. Yu, and D.H. McLaughlin, "A Study of the Administration of the 1994 Trial State Assessment," in *Quality and Utility: The 1994 Trial State Assessment in Reading, Background Studies* (Stanford, CA: The National Academy of Education, forthcoming).

Further evidence for the relationship between the size of the TSA and school recruitment problems is provided by the just completed experience of the 1996 state NAEP. Although the Panel has not collected systematic data on the 1996 assessment, increases in school complaints about participation and/or difficulties with recruitment have been anecdotally reported by several state assessment directors, including at least one in a state that was nevertheless able to deliver 100 percent school participation because of state mandate.[9] Problems seem most acute in the smallest jurisdictions, in which the usual sample size for one subject at one grade (2,500) may represent 30 percent or more of the total number of students in that grade. Matters are clearly compounded when more than one subject per grade is assessed, requiring the recruitment of a very high percentage of available schools in every assessment cycle to meet sampling requirements.

*In summary, the Panel finds that participation rates for originally sampled public schools in the 1994 TSA generally ranged from acceptable to good, although rates varied considerably across jurisdictions.*

Despite these generally positive findings, there are indications that the increased numbers of schools and students required when multiple subjects and grades are assessed cause a burden that could threaten state NAEP participation. The Panel recommends that NCES and NAGB pay close attention to trends in school participation rates. Additionally, they should consider design changes that could decrease sample size requirements or otherwise reduce burden without compromising the overall quality of the assessment. Applicable design changes could include relatively circumscribed modifications such as applying the principles of finite sampling to create a different set of rules for the smallest states. Reduced respondent burden could also be affected as one outcome of a more radical redesign of NAEP, and various versions of the latter are currently being debated by NAGB and other interested parties. Additional discussion of respondent burden and its potential impact on participation will be provided in the Panel's forthcoming capstone report.

## *Impact of School Nonparticipation*

Following the previous TSAs, the Panel noted that states with lower before-substitution school participation rates tended to have higher average achievement scores. As shown in table 3.4, the negative correlation between initial participation rates and average performance scores was -.286 in 1990. In 1992, the correlations were -.447 for grade-four reading, -.367 for grade-four mathematics, and -.361 for grade-eight mathematics.[10] In 1994, the correlation was again negative: -.302 for grade-four reading (based on public schools only). Despite the fact that only the correlations for 1992 were statistically significant, this consistent negative relationship

[9] E. Hartka and F. Stancavage, "Perspectives on the Impact of the Trial State Assessments: State Assessment Directors, State Mathematics Specialists, and State Reading Specialists," in *Quality and Utility: The 1994 Trial State Assessment in Reading, Background Studies* (Stanford, CA: The National Academy of Education, forthcoming).

10 The National Academy of Education, *The Trial State Assessment: Prospects and Realities* (Stanford, CA: Author, 1993), 21.

over all TSAs raised concerns that some states' high average performance scores might be caused by the nonparticipation (in those states) of schools whose students would have scored lower than the state average.

**Table 3.4.** ***Correlations between weighted before-substitution public school participation rates and average state proficiency for public school students***

| 1990 TSA | 1992 TSA | | | 1994 TSA |
|---|---|---|---|---|
| Grade-8 mathematics | Grade-4 reading | Grade-4 mathematics | Grade-8 mathematics | Grade-4 reading |
| -.286 (n = 40) | -.447* (n = 43)** | -.367* (n = 43)** | -.361 (n = 44) | -.302 (n = 44) |

*Statistically significant at $p < .02$

**Does not include the Virgin Islands, which exercised its option not to release fourth-grade results in 1992

In theory, refusing and substitute schools should not differ significantly provided that refusing schools are replaced with similar schools. As discussed however, there is a strong likelihood that refusing schools will differ from participating schools in ways that are not controlled by matching substitutes on a limited number of variables. If student achievement is affected by any of these uncontrolled differences, nonresponse bias will occur when even a small percentage of originally sampled schools refuse to participate.

The Panel therefore conducted, as part of its 1994 evaluation, a study to further investigate the relationship between school nonparticipation and other school characteristics, particularly student achievement.[11] Eleven states were selected for study.[12] All had state reading assessments in grades three, four, or five, and all but one had low initial participation rates in either the 1992 or 1994 TSA. Recent state reading scores were obtained for all of the originally sampled schools and all of the substitute schools in the 11 states, and the relationship between school-average TSA scores and school-average state assessment scores was examined. Because different states use different reading assessments, the strength of this relationship differed from state to state. In seven of the study states however, the relationship was at least moderate (correlations across schools ranged from .45 to .72) and adequate for the purposes of the study.

State assessment scores were used to predict average TSA scores for each school, and originally participating, refusing, and substitute schools were compared on the basis of

[11] E. Hartka, M. Perie, and D.H. McLaughlin, "Public School Nonparticipation Study," in *Quality and Utility: The 1994 Trial State Assessment in Reading, Background Studies* (Stanford, CA: The National Academy of Education, forthcoming).

[12] The 11 study states were California, Connecticut, Indiana, Maine, Massachusetts, Michigan, Montana, Pennsylvania, Rhode Island, Tennessee, and Wisconsin.

both state assessment scores and predicted TSA scores.[13] *Overall, the analyses provided no evidence that the pattern of school refusals and substitutions raised the mean state TSA scores in any of the study states. This conclusion is reassuring and suggests that current substitution and nonresponse adjustment procedures are adequate and that the 70 and 85 percent cutoffs used in the NCES reporting standards are reasonable, or at least not too liberal. However, the study only involved a limited number of states. In most cases, the numbers of refusing and substitute schools was small, and the negative correlation between average state scores and before-substitution rates was not explained by the findings.* Further investigation of the impact of nonresponse, based on a larger sample of states, is therefore warranted. Such a study is in progress, and results should be available in late 1996.[14]

In further analyses, demographic characteristics of the schools and their communities (e.g., percent minority enrollment) were added from other NCES data sources and examined in relation to participation status. Some patterns were observed. In the nine study states with both refusing and substitute schools, refusing schools tended to be larger and have higher percentages of minority students than their substitutes. Similar relationships had been observed in the 1992 TSA, when refusing schools had evidenced somewhat lower median incomes and more minority students than substitute schools. Combined, this evidence suggests that schools with higher minority enrollments are more reluctant to take on the burden of participation, perhaps because high minority enrollment tends to be associated with fewer school resources and greater student needs. If this is true, increasing school participation rates might require that TSA results be made more valuable to individual schools (something that is unlikely under the present NAEP design) or that schools be provided with more assistance to facilitate their participation.

## *Sampling and Participation of Students*

In each selected public school, the NAEP contractor drew a systematic sample of about 30 students from a list of all grade-eligible children that the school had provided about two-and-a-half months before the assessments were to begin.[15] Between the time the original lists were drawn and the time of assessment, some students left their schools and others entered. Those who had withdrawn were necessarily lost from the sample, but a compensatory procedure was used to allow other students changing schools during this period to be included in the sample. Specifically, each school was asked to maintain a supplementary list of all students who had moved *into* the school after the previous grade-eligible list was sent to the contractor. Assessment

[13] Among the 11 study states, the numbers of originally particpating schools ranged from 65 to 105 in 1994 with a median of 86. Refusing schools ranged from 3 to 38 (median = 21), whereas the numbers of substitute schools ranged from zero to 17 (median = 6).

[14] A further study of the impact of school nonparticipation in a larger group of states is currently being conducted by Bruce Spencer, Northwestern University.

[15] In the smallest jursidictions however, the number of students per schools was much higher. For example, in Delaware, up to 120 students per school were selected. The sampling of students in DODEA schools also differed in some details; see Westat (1995), 3-46.

administrators were given guidelines for sampling from these supplemental lists, and selected students were added to the school sample.

Table 3.5 shows that, in 1994, 3.8 percent of all originally sampled students had withdrawn by the date of the assessment, and 3.0 percent had been added through supplemental sampling. Such discrepancies between the numbers withdrawn and added are typical of previous TSAs and are presumably caused by temporary or permanent withdrawals from public school that are not offset by students newly entering the public school system during this time.

**Table 3.5.** ***Percentages of students withdrawn prior to assessment, added from supplemental sample lists, absent from original assessment sessions, and subsequently assessed in make-up sessions: 1994 TSA***

| | |
|---|---|
| Withdrawn before assessment* | 3.8 |
| Added from supplemental sample* | 3.0 |
| Absent from original assessment session** | 5.3 |
| Assessed in make-up session** | 0.8 |

* Percent of all originally sampled students

**Percent of all eligible sampled students (excludes withdrawn and excluded students)

SOURCE: B.D. Spencer, "School and Student Sampling in the 1994 TSA: An Evaluation," in *Quality and Utility: The 1994 Trial State Assessment in Reading, Background Studies* (Stanford, CA: The National Academy of Education, forthcoming).

NOTE: Data in this table are unweighted.

In addition to student withdrawals, the final student samples in each school are affected by the exclusion of students with disabilities or limited English proficiency (discussed in chapter 4), and the nonparticipation of eligible students. Overall, the participation rates for eligible public school students were very high in 1994 (as they also were in previous TSAs), exceeding 94 percent in every state. The primary reason for nonparticipation of eligible students was absence; other, far less frequent reasons included parent or student refusal. When three or more students were absent from the original assessment session, schools were required to hold make-up sessions. As can be seen in table 3.5 above, the proportion of absent students recovered through make-up sessions was small. ***Nevertheless, because some schools may be motivated to exclude low performing or disruptive students, and because a certain amount of effort is in any case required to gather all of the sampled students (who may be spread through several classrooms), the requirement for make-up sessions seems useful for motivating maximum participation at originally scheduled sessions. The Panel suggests that NCES continue to require make-up sessions when three or more students are absent.***

## *Impact of Under Sampling and Nonparticipation on Student Samples*

*As noted, participation of sampled students in the TSAs has been uniformly high across states and, consequently, has contributed relatively little to the potential for nonresponse bias—particularly in comparison to the much more varied school nonparticipation rates.* Nevertheless, analyses of TSA and other assessment data have shown that the kinds of students lost to the assessment through high mobility or absence tend to score lower than average. Consequently, there is some potential for overestimating student achievement in jurisdictions where high mobility and frequent student absences are common. This might become more of a concern if NAEP begins to report results for smaller populations of students, such as individual districts.

## *Weighting*

In a survey as complex as the TSA, sampling weights are required to make valid inferences from the student samples to the respective populations from which they were drawn. Participants are selected to represent each component of the total sample, and their responses must be weighted more or less heavily in the ensuing analyses, depending upon the numbers of other students (who were not selected or did not participate) that each represents. The weighting procedures used in the 1994 TSA were comparable to those used in previous TSA assessments. First, base weights were established that reflected the consequences of the sampling design by incorporating the probability of selecting a school and the student within a school. After the assessment, the base weights were adjusted for two sources of nonparticipation: school level and student level. This is a very important component of the weighting process because, as noted, schools and students that refuse (or are unable) to participate are likely to differ in systematic, but unknown, ways from those who do participate. To minimize the effects of nonparticipation, the weights of the nonparticipants must be redistributed across those with similar characteristics who did participate.

In 1994, adjustments for nonparticipation of schools were performed separately for each state, and took account of school-level characteristics, such as type of school and size of community, which were also considered when the original sample was drawn. Weights for nonparticipating students within participating schools, however, were adjusted primarily on the basis of these same school-level characteristics, with only one student-level characteristic—a dichotomous age variable—contributing to the reassignment of weights. The adjustments would probably have been more effective had additional student-level variables such as gender, ethnicity, or Title I status been considered because participating and nonparticipating students—even within the same school—tend to differ on many characteristics that are also associated with differences in overall test performance.

*In summary, with regard to the 1994 TSA, as with previous TSAs, the Panel found that the weighting procedures used to compensate for different selection rates among different categories of students was reasonably straightforward and that the adjustments for public school nonparticipation were appropriate. The adjustments for student nonparticipation, however, omitted the effects of student-level variables that could have further reduced bias by accounting for more of the within-school variation in student performance.* ***The Panel therefore suggests, as it has in both its previous TSA evaluation reports, that NCES and the NAEP contractor explore a more precise correction for student nonresponse based on the demographic and program status variables that are already being collected for all participating and nonparticipating students (e.g., their sex, race, and chapter I status).***

## *Nonpublic School Samples*

### *The Panel's 1992 Recommendation to Include Nonpublic School Students*

NAEP assessment results are generally taken to be representative of the achievements of all U.S. students. Representation is not complete, however, because certain classes of students are excluded from NAEP results by design. Prior to 1994, private school students typically made up the largest excluded group in each state. It is estimated that approximately 11 percent of the nation's fourth-grade students are enrolled in private schools, but the percentage varies widely across states. In its first report, *Assessing Student Achievement in the States,* the Panel noted that "state NAEP data would better reflect education achievement and make state results more readily comparable if results for all students, private as well as public were included. Ultimately, because of the significant variation from state to state in private school enrollment, the exclusion of private school students from the TSA will diminish its utility." The Panel went on to recommend that "future authorizations for state NAEP include adequate resources to sample private school students, thereby increasing the comparability of results from one state to another."[16]

By 1993, when the Panel released *The Trial State Assessment: Prospects and Realities,* Congress had acted on this recommendation and authorized inclusion of nonpublic school students in the 1994 TSA.[17] The Panel further affirmed the value of including these and other excluded groups in formulating its *inclusiveness principle,* which states that "to the degree technically, ethically, and financially possible, NAEP should assess an inclusive sample of all children whose ages would ordinarily place them in

16 The National Academy of Education, *Assessing Student Achievement in the States* (Stanford, CA: Author, 1992), 12.

17 As noted earlier in the chapter, the category of TSA schools that include private schools also encompasses small numbers of Bureau of Indian Affairs schools and domestic Department of Defense schools in some jurisdictions. Consequently, the category is referred to as "nonpublic schools" here and elsewhere in the report.

the 4th, 8th, or 12th grade."[18] The Panel's intention was to enhance the utility of overall state scores by including nonpublic school students and other groups excluded by the current design. The Panel did not intend, nor expect, separate reporting of state nonpublic school results.

## *Sampling of Schools*

The sample design for nonpublic schools in the 1994 TSA was based on the assumption that the results would be aggregated with public schools for reporting. For this purpose, the proportion of nonpublic schools in each state's sample needed only to be as large as the proportion of nonpublic schools in the state. Consequently, the samples of nonpublic schools drawn for participating states averaged about 15 schools, and six states had samples smaller than 10.

However, once the process of including nonpublic schools had begun, it became apparent that there were rather compelling reasons for separate reporting of public and nonpublic school results. Separate reporting of public school results was necessary (and always intended) to provide appropriate comparisons to the first two TSAs, which did not include nonpublic schools. Moreover, the state officials that are responsible for implementing the TSAs are primarily interested in public school results because these are the schools that fall under their jurisdiction and responsibility.

The intention to provide separate reporting of public school results generated constituent pressures for separate reporting of nonpublic school results as well. In particular, it became obvious, once the matter was considered, that nonpublic schools would have little incentive for participating if they were not offered recompense in the form of useful information about the performance of their own students.

Unfortunately, the convergence of these circumstances led to a situation in which NCES reported nonpublic school results for the 1994 TSA based on samples not well suited to the task. Nonpublic schools are very diverse and, as noted, the nonpublic school samples were quite small. Consequently, even if every one of the sampled nonpublic schools had participated, the resultant estimates for nonpublic school achievement would have had such wide bands of uncertainty as to be uninformative and potentially misleading.

The data in the *NAEP 1994 Reading Report Card for the Nation and the States* illustrates the implications of these wide bands of uncertainty. Because NAEP is based on a sample, the values reported are estimates, and information on the standard errors of estimate for these values are given. Footnotes to each table explain that "it can be said with 95 percent certainty that for each population of interest, the value for the whole population is within plus or minus two standard errors of the estimate for the

[18] The National Academy of Education, *The Trial State Assessment: Prospects and Realities* (Stanford, CA: Author, 1993), 96. As presented in chapter 1 of the current report, this principle has been slightly revised to remove the reference to specific grade cohorts. The choice of assessment cohorts is a policy decision not related to inclusion.

sample." Thus, for example, the estimated average reading proficiency for public school students in Iowa is 223 on the NAEP scale.[19] (See table 3.6.) The standard error of estimate is given as 1.3, meaning that the true average reading proficiency for Iowa public school students is almost certain to lie between 220.4 and 225.6. For nonpublic school students in Iowa however, estimated average reading proficiency is 232, with an estimated standard error of estimate of 4.2. Therefore, the confidence interval for the nonpublic school estimate lies in the 16-point range between 223.6 and 240.4, and it is *six-and-a-half times as large* as the confidence interval for the public school estimate.

The Iowa standard errors of estimate are typical of the estimates for state-average proficiency scores and are based entirely on the size and composition of the sample of schools, not on the percent of schools responding. In fact, Iowa had a 100 percent weighted before-substitution response rate for the 16 nonpublic schools in its sample, but only an 85 percent weighted before-substitution response rate for the 108 public schools. After substitution, the weighted school participation rates were very high for both types of schools.[20]

**Table 3.6.** ***Estimated average 1994 reading proficiencies for Iowa public and nonpublic school students, school sample sizes, and weighted school participation rates***

| | Public Schools | Nonpublic Schools |
|---|---|---|
| Estimated average reading proficiency (standard error of estimate in parentheses) | 223 (1.3) | 232 (4.2) |
| Confidence interval for estimated average reading proficiency | 220.4 - 225.6 | 223.6 - 240.4 |
| Number of eligible schools in original school sample | 108 | 16 |
| Weighted before-substitution school participation rate | 85% | 100% |
| Weighted after-substitution school participation rate | 99% | 100% |

SOURCES: J.R. Campbell, P.L. Donahue, C.M. Reese, and G.W. Phillips, *NAEP 1994 Reading Report Card for the Nation and the States* (Washington, D.C.: National Center for Education Statistics, January 1996), 34, 105-106, and J. Mazzeo, N.L. Allen, and D.L. Kline, *Technical Report of the NAEP 1994 Trial State Assessment Program in Reading* (Washington, D.C.: National Center for Education Statistics, December 1995), 65-66.

[19] J.R. Campbell, P.L. Donahue, C.M. Reese, and G.W. Phillips, *NAEP 1994 Reading Report Card for the Nation and the States* (Washington, D.C.: National Center for Education Statistics, January 1996), 34.

[20] The weighted after-substitution response rate for Iowa public schools was 99.

In order to obtain more precise state-level estimates for nonpublic school students, much larger samples of nonpublic schools would be needed. In states with few nonpublic schools, these samples could be extremely burdensome because the same nonpublic schools would have to agree to participate repeatedly and for every grade and subject included in the state assessments.

## *School Participation Rates*

The weighted before-substitution participation rates for nonpublic schools were much lower than those for public schools (the average state rates were 64 percent for nonpublic schools compared to 90 percent for public schools), and much more variable (ranging from 0 percent in three states to 100 percent in others). Altogether, 29 states had initial nonpublic school participation rates below 85 percent; 18 of these fell below the 70 percent standard required for reporting. There were a number of reasons for these lower rates. First, because the total numbers of nonpublic schools in the samples for most states were small, only a few schools had to refuse before participation rates fell to unacceptable levels. Second, some nonpublic schools wanted to participate but, more frequently than public schools, had difficulties getting permission from their governing bodies or lacked the resources necessary for participation (e.g., an extra staff person who could be spared for the one-day training required of assessment administrators or even an extra classroom in which to hold the assessment). Third, the recruitment process for nonpublic schools was problematic in many states. The U.S. Department of Education had decided that, under the NAEP authorization for 1994, the federal government could not fund recruitment for the TSAs, even for nonpublic schools. However, the state education departments that handled the recruitment for public schools were not necessarily well placed to recruit nonpublic schools. Whereas some state departments of education do exercise oversight responsibility over nonpublic schools or enjoy a cooperative association with them, many others have only limited contact with these schools or are constrained by statute from interfering with and/or spending any money for the benefit of nonpublic schools.[21]

Not only was nonparticipation higher on average among nonpublic schools, but the pattern of nonparticipation was almost certainly nonrandom. In general, the larger, more established nonpublic schools were most likely to participate, whereas the smaller, newer, and less traditional schools were less likely to do so. In some cases, these refusals reflected the principles of groups that had turned to nonpublic schooling in order to avoid government regulation or influence; in other cases, very new or very small schools were disproportionately likely to lack the internal resources or external support systems that would make it possible for them to bear the in-kind costs of participation.[22] Because there is so much diversity among nonpublic schools,

[21] P.J. Devito, "The Future of NAEP from the States' Perspective," in *Quality and Utility: The 1994 Trial State Assessment in Reading, Background Studies* (Stanford, CA: The National Academy of Education, forthcoming).

[22] These same smaller, newer, and less traditional nonpublic schools were also the most likely to be accidently omitted from the school lists used in sampling, further exacerbating the under-representation among this class of schools.

and many of the relevant variables were not well documented on the available school lists, there was also less opportunity to reduce nonresponse bias through careful matching of substitute schools or finely-tuned weight adjustments.

*In summary, the Panel concludes that the nonpublic school samples used in 1994 did not support meaningful comparisons between public and nonpublic schools, even in jurisdictions where nonpublic school response rates were high. Furthermore, there was a high percentage of jurisdictions in which the nonpublic school response rates fell below the participation standards that NCES requires for reporting, and it appears likely that uncorrected nonresponse bias diminished the generalizability of these results in other jurisdictions as well.* The recently completed 1996 state NAEP assessments used a similar nonpublic school sampling plan that the Panel judges to be again unsuitable for separate reporting, despite the fact that the minimum sample size was raised slightly from six to 10 schools. Efforts were made, in the 1996 state assessments, to improve response rates by allowing the NAEP contractor to take charge of nonpublic school recruitment. Preliminary evidence suggests, however, that these efforts were not as successful in raising participation rates as had been hoped, and that the state rates remain lower than the nonpublic school participation rates obtained in the national sample. (In national NAEP, both school recruitment and session administration are the responsibility of the contractor. Consequently, national cooperation rates may be enhanced by the fact that participating schools do not have to commit as much staff time.)

Finally, the Panel has serious reservations about the separate reporting of state-level nonpublic school results that go beyond issues of sampling and participation rates. *Because of the many powerful influences affecting student achievement, NAEP results, by themselves, are not a sufficient basis for comparing the quality of educational programs. Nevertheless, differences in student performance are frequently, and simplistically, attributed to differences in type of school without due consideration of differences in the populations of students that they serve. The separate reporting of nonpublic school results therefore holds considerable potential for incorrect inferences about differences in performance between public and nonpublic schools.*

The Panel recommends that NAGB and NCES stop separate reporting of state-level nonpublic school results, but, where participation rates are sufficiently high, continue reporting state-level results for public and nonpublic schools combined and for public schools only. Furthermore, the reports should include prominent warnings about the invalidity of simplistic comparisons between public and nonpublic schools in order to discourage efforts to derive such comparisons by subtracting public school means from the combined public and nonpublic school results. These warnings should be illustrated by concrete examples to underscore their significance.[23]

[23] A counter argument is that separate reporting of nonpublic school results, such as those obtained in 1994, at least serves to make explicit the lack of precision (wide confidence bands) associated with these estimates. In the absence of such information, the reader is left to make his or her own estimate of student performance in nonpublic schools by subtracting the published mean proficiency for public school students from the published mean proficiency for all students. On balance however, the Panel believes that it is more problematic to "legitimate" the nonpublic school results by reporting estimates based on inappropriate samples. If these results must be reported, then the Panel suggests that they be reported *only* as ranges and not as point estimates. (For Iowa, for example, this would mean reporting that average proficiency for nonpublic school students falls between 223.6 and 225.6 but *not* reporting the associated point estimate of 232. See discussion on pp. 42-43) As noted in the Panel's recommendations, using larger school samples and reducing nonresponse bias would also go further towards producing defensible nonpublic school results.

At the same time, NAGB and NCES should explore alternative strategies (other than separate state-level reporting) for motivating the participation of nonpublic schools. One proposed course of action would be to offer more detailed reporting of nonpublic school results at the *national* level by basing the analyses on aggregated data from the national and state samples of nonpublic schools. The latter appears to be a plausible alternative in view of the fact that nonpublic schools, in any event, may be less interested in state-level comparisons to public schools than in the more detailed comparisons among different types of nonpublic schools that the larger, combined state and national samples would permit.

If NAGB and NCES nevertheless wish to continue reporting state-level nonpublic school results, then larger samples of nonpublic schools must be used. In the latter case, NCES and the NAEP contractors must also undertake studies to estimate the extent of bias introduced by the fact that different types of nonpublic schools tend to participate at different rates, and to find ways to reduce this bias.

## *The Administration of the 1994 TSA*

The administration procedures for the 1994 TSA were similar to those used in previous TSAs. Underlying the procedures is the fact that, by law, responsibility for the administration is shared between the NAEP contractor and state or local personnel. This is in contrast to the national assessment, wherein the contractor has primary responsibility for all aspects of administration. The actual administration of the TSA, in addition to certain related activities, falls under the purview of local assessment administrators who are, in most cases, teachers or staff members recruited from the sampled schools. Contractor staff provide training for the assessment administrators, contact them shortly before the scheduled date of the assessment to confirm receipt of materials and other assessment arrangements, and travel to a sample of schools on the day of the assessment to monitor the session administration.

In its evaluation of the administration of the 1990 TSA, the Panel concluded that the assessment was carefully planned and implemented and that the training of assessment administrators was generally successful. No statistically significant differences in student performance were observed between subjects in monitored sessions and sessions that were not monitored. However, the Panel found evidence that several types of administration irregularities, when they occurred, were significantly associated with lower average performance scores on the TSA. In addition, TSA subjects performed slightly, but significantly better than a matched sample of national NAEP subjects.

With respect to the 1992 TSA, findings paralleled those for 1990. The Panel concluded that the overall quality of the administration was very high. Administration irregularities, though infrequent, were associated with lower average performance. In addition, the Panel noted that significant numbers of fourth graders (who were included in the TSA for the first time in 1992) experienced difficulties interpreting and providing meaningful responses to the background questions that cover demographic and home background characteristics. As in 1990, no significant differences in student

performance were observed between monitored and unmonitored sessions. Comparisons between the TSA and a matched sample from the national assessment showed the TSA subjects doing slightly better on average, but, unlike 1990, these differences were not statistically significant.

In 1994, the Panel again compared the performance of students in state and national samples and in monitored and unmonitored sessions, and examined the frequency of session irregularities and the relationship of such irregularities to student performance. With regard to the latter effort, the statistical analyses used in previous evaluation cycles were supplemented with direct observations of more than 50 administration sessions, in 13 states.

*In general, the findings for the 1994 administration study indicate that the design of the TSA administration was good and the overall quality of the implementation was high.* Specific details from the study are discussed in the following sections.

## *Comparisons of TSA and National Performance*

In evaluating the TSAs, consideration of the consistency and comparability of the state and national assessments is crucial. A major benefit of the state NAEP program is that it provides states with uniform student data that are directly comparable to results from other states and the nation. A strong test of this consistency and, by extension, the adequacy of the state assessment administration conditions, is to compare overall performance for each grade and subject of the TSA against the overall performance of a matched sample drawn from national NAEP.[24]

*In 1990, the state-to-national comparison for eighth-grade mathematics found a small but statistically significant performance difference favoring students in the state samples. Similar performance patterns were found in both 1992 and 1994, but the differences between the two samples were not statistically significant.*[25] It is possible that the slight but persistent performance difference favoring the state samples reflects higher levels of student motivation related to taking the assessment under the supervision of someone that the students know (the national assessment, as noted, is administered by NAEP contractor staff) or to the communicated enthusiasm of school staff who place a greater value on performing well when results are to be reported at the state level.

## *Comparison of Monitored and Unmonitored TSA Sessions*

In 1990, employees of Westat, the NAEP contractor for administration, monitored 50 percent of the TSA sessions in every jurisdiction. Schools did not know in advance

[24] A matched sample must be drawn for comparison because national NAEP is designed to produce estimates for the entire nation and not just for the aggregation of states that participated in the TSA.

[25] E. Hartka and D.H. McLaughlin, "A Study of the Administration of the 1992 National Assessment of Educational Progress Trial State Assessment," in *The Trial State Assessment: Prospects and Realities: Background Studies* (Stanford, CA: The National Academy of Education, 1994) and E. Hartka, J. Yu, and D.H. McLaughlin, op. cit.

whether or not they would be monitored. The intent was to collect information on the adequacy of the administration and, if necessary, to intervene in order to prevent major violations of protocol. In this way, NCES and the NAEP contractors were assured of having sufficient data on which to report assessment results, even if some or all of the data from unmonitored sessions proved unusable. For 1992, and consistent with the Panel's recommendation, the 50 percent monitoring rate was maintained despite the fact that no major problems had been observed in 1990. After the experience of two successful administrations however, a lower monitoring rate seemed acceptable for 1994—at least in jurisdictions that had participated before.

In 1994 therefore, the monitoring rate was reduced to 25 percent except for nonpublic schools and jurisdictions that were participating for the first time. Because nonpublic schools were over-represented in the sample of monitored schools, and because nonpublic school students perform substantially better than public school students on average, the Panel analyzed the monitored/unmonitored comparisons separately by type of school. *For both public and nonpublic schools, students in monitored sessions performed slightly better than students in unmonitored sessions, replicating the pattern observed in previous TSAs. In neither case were the differences between monitored and unmonitored sessions statistically significant. In the technical manual for the 1994 TSA however, the Educational Testing Service (ETS) reported that, for nonpublic schools in a small number of jurisdictions, students in unmonitored sessions performed noticeably better than students in monitored sessions.* This finding is potentially troubling because it could suggest that significant improprieties (e.g., allowing students more time, helping students with their answers) were occurring in the unmonitored schools. ETS found no further evidence for such deviations however, and concluded that, in each of the affected jurisdictions, the number of nonpublic schools was so small that the comparisons could have been heavily affected by the particular nonpublic schools that were assigned to each condition, even though proper random sampling techniques had been followed in making these assignments.[26]

In light of these findings, the Panel would have recommended that the 1994 monitoring rate of 50 percent for nonpublic schools be left in place for the time being, and that the relationship between monitoring status and student performance continue to be tracked. However, the nonpublic school monitoring rate was dropped to 25 percent for 1996, and this, combined with the already small size of the nonpublic school samples, will make it nearly impossible to repeat the within-state comparisons carried out in 1994.

The Panel recommends, however, that a low level of random monitoring be maintained indefinitely for both public and nonpublic schools. Such monitoring may play a role in motivating more thorough assessment administrator preparation and in deterring infractions of the administration guidelines.

[26] J. Mazzeo, N.L. Allen, and D.L. Kline, *Technical Report of the NAEP 1994 Trial State Assessment Program in Reading* (Washington, D.C.: National Center for Education Statistics), 373-392.

In each TSA, the Westat quality control monitors completed rating forms on which they noted the occurrence of any irregularities or disruptive conditions in the sessions they observed and evaluated the overall adequacy of the assessment administrator's performance. For 1994, the Panel once again reviewed these data and examined the relationships between characteristics of the administration and student performance. In addition, the Panel sent its own observers to 56 schools in 13 states to independently monitor the administration of the assessment. The latter sessions comprised some that were also monitored by Westat and some that were not, and included disproportionate numbers of nonpublic schools and schools with lower estimated family incomes. Nonpublic schools were of interest because this was the fist time they were included in the TSA. Low income schools were over sampled because previous analyses suggested that session irregularities or adverse administration conditions might be expected to occur with greater than average frequency in these schools.

Panel observers attended Westat training sessions to familiarize themselves with administration guidelines, then completed observation forms for each session that covered many of the same aspects of the administration as the Westat rating forms.

Although Panel observers were somewhat more critical of assessment administrator performance than Westat's monitors, they still found a reasonable level of compliance with the established protocol for conducting the assessment. Most observers reported that the assessment administrators were careful to follow procedures and that students were well behaved and appeared to work hard on the assessment. It was observed generally that students appeared less restless and worked harder where schools prepared in advance and assessment administrators were familiar with the procedures, followed them closely, and encouraged students to stay on task.

In some cases however, Panel observers witnessed various departures from scripted procedures and sessions that did not proceed smoothly. The main problems fell into a few areas, the most troublesome of which concerned the portion of the administration devoted to answering the student background questions. As noted in 1992, fourth-grade students often seemed confused by these background questions (race/ethnicity, parent's education, reduced price lunches, etc.) and had trouble coding their school identification numbers and home zip codes onto machine scannable grids. When students asked questions about these and other items, Panel observers reported that almost half of the assessment administrators were not clear about how much help to give during this portion of the assessment. Perhaps, because the background items come at the very beginning of the assessment, the assessment administrators were concerned about getting off schedule or with setting a precedent for "help giving" that would be awkward to back away from in the later, cognitive, portions of the assessment. In fact however, assessment administrators were instructed to answer student questions and to help with the interpretation of the items in the background section, particularly because these questions yield very important analysis variables that are not currently available from any other source. Furthermore, it has been repeatedly shown that the reliability of these background data is disappointingly low for fourth-grade students, at least in part because some of the students do not understand the questions.

Using the data from all monitored sessions, average performance in classrooms with administration irregularities was compared to average performance in classrooms with no irregularities. After adjusting for race and gender, the comparison found average performance scores to be significantly higher when any of the following conditions prevailed: 1) the supplemental student sampling was completed on time; 2) the Teacher Questionnaires were distributed on time; 3) the script for "Directions" was read without error (verbatim); 4) the script for "Section 1: Background" was read without error; 5) the Westat monitor's overall evaluation was that the assessment administrator functioned competently; 6) students were cooperative; 7) students appeared to have had positive attitudes toward the assessment; and 8) the assessment administrator was a principal, vice-principal, or superintendent. The same analysis following the 1992 TSA yielded similar results: higher student performance was then significantly linked to verbatim reading of the instructions, competent assessment administrators, positive student attitudes and cooperation, and administration by principals and superintendents.[27]

The reader is cautioned that, in any correlational study, it is not possible to attribute causality with certainty. Nevertheless, it seems likely that some or all of the factors identified may have enhanced performance by increasing student motivation. In low-stakes assessments like NAEP, which have no consequences for individual students, it is generally assumed that levels of student performance are somewhat dampened by low student motivation. According to the Panel's observers, administrators in sessions that yielded higher average performance read the script accurately, effectively helped students through difficult procedures such as coding their student identification numbers onto the booklet covers, and kept students on task throughout the assessment. The latter involved making the students aware that their cooperation and effort was expected, walking around during the assessment to re-enforce the perception that the students' behavior was being noted, and specifically redirecting students whose attention had obviously wandered. Moreover, students tended to be well behaved for principals and superintendents and took the test more seriously when school staff treated the TSA as an important activity.

*The Panel concludes that, in 1994, as in previous TSAs, most sessions ran smoothly and were well administered in both nonpublic and public schools. However, the Panel again observed, as it did in 1992, that fourth-grade students had difficulty completing the background questions, thus compromising the quality of the socioeconomic measures derived from these data.* The Panel was therefore pleased to learn that NCES has undertaken studies to identify the kinds of background information that fourth graders can reliably provide and to examine alternatives for obtaining data on other important background variables.

## *Summary*

In summary, sampling and administration procedures for the 1994 TSA appear to have been well designed and successfully executed, at least in the case of public schools. Public school samples were selected appropriately, school recruitment proceeded smoothly, and school participation rates were acceptably high in most jurisdictions.

[27] E. Hartka, J. Yu, and D.H. McLaughlin, op. cit.

Anecdotal evidence from the 1996 assessment, however, suggests that school recruiters in some states have been experiencing greater local resistance to participation, particularly in smaller states where the sampling burden of assessing multiple subjects and grades was felt most acutely.

Within participating schools, student participation rates were also quite high, and adjustments for nonresponse were done competently. Two minor problems were the use of a procedure for recruiting substitute schools that could have, but apparently did not, lead to unnecessary bias, and the failure to consider student-level variables in redistributing the weights of nonresponding students. On a positive note, the Panel's study of the impact of public school nonresponse, although restricted to only 11 states, produced no evidence that the pattern of school refusals and substitutions raised mean state TSA scores compared to the scores that would have been obtained if all originally sampled schools had participated.

For nonpublic schools however, the Panel concluded that the school samples, which were drawn proportional to the numbers of nonpublic schools of each state, were too small to produce results of the accuracy needed to support separate nonpublic school reporting. Furthermore, nonpublic school participation rates were extremely low in many of the states, and there were almost certainly nonrandom differences between the types of nonpublic schools that did or did not choose to participate. The latter circumstance adds bias, in addition to uncertainty, to the resultant estimates of nonpublic school student achievement. When the recommendation to include private schools in the TSA was made by the Panel and accepted by Congress and NCES, none foresaw the pressures that would arise for separate reporting by school type nor fully considered the implications of such reporting. Consequently, the Panel holds that conditions for nonpublic school participation must be reconsidered now and separate reporting stopped or nonpublic schools samples substantially enlarged.

Finally, the administration of the 1994 TSA appears to have been conducted appropriately in almost all cases. As in previous years however, there was evidence that irregularities in administration, when they did occur, were associated with somewhat lower student achievement. In particular, assessment administrators who were not well organized in advance of the assessment, who had difficulty reading the script without errors, and who failed to monitor the students in ways that kept them appropriately on tasks may have had a depressing effect on student scores. A more widespread problem in session administration was the fact that fourth-grade students frequently evidenced problems responding to the background questions that currently provide the only basis for important socioeconomic variables in NAEP. This is a known problem, and NCES has undertaken a series of studies that will hopefully shed additional light on the functioning of the student background questions as well as identify other potential sources for socioeconomic variables that can be used in NAEP analyses.

# 4 The Assessment of Students with Disabilities or Limited English Proficiency

## Introduction

Fundamental to NAEP's role as a key indicator of what the nation's students know and can do is the concept that NAEP results, insofar as possible, should represent *all* U.S. students. Similarly, results for each state should be representative of all students in that state and, to the extent that some students are nevertheless excluded because they cannot be appropriately assessed using the NAEP instruments and procedures, the kinds of students excluded from each state should be comparable. These issues, combined with the observation that different states excluded different percentages of their student populations in the two previous TSAs, led the Panel to include special studies of assessability and exclusion as part of its evaluation of the 1994 TSA.

Furthermore, recent education trends—including the passage in 1994 of Public Law 103-227, the *Goals 2000: Educate America Act,* which calls for academic standards and assessments that are meaningful, challenging, and appropriate for all students—have prompted NCES and NAGB to look more closely at NAEP procedures regarding students with disabilities (IEP students)[1] or limited English proficiency (LEP students). Specifically, NCES and NAGB have expressed their commitment to increasing inclusion of such students in the NAEP assessments, while also acknowledging the need to maintain high technical standards, continuity of trend data, and a balance of resources across important programmatic goals. The new exclusion guidelines and testing accommodations developed for the 1996 assessment reflect these commitments. The Panel, which in its own writings has promulgated a set of guiding principles that embrace both *inclusiveness* and *quality,*[2] concurs with NAEP's overall goals for IEP and LEP students and designed its 1994 studies to address the bounds of appropriate assessment as well as the comparability of exclusion decisions across states.

The Panel's studies, which are described in this chapter, involved site visits to 188 schools in seven states. Site visitors collected data from IEP and LEP students who had been selected for the 1994 TSA, from those students' teachers, and from the local personnel who administered the TSA assessments in their schools. The chapter begins with background information on the IEP/LEP exclusion procedures in effect through 1994 and the exclusion rates observed in each of the three TSAs. This is followed by a section on IEP students that presents information from the NAEP IEP/LEP

[1] Students with disabilities were called IEP students in the nomenclature of the 1994 NAEP because of the individualized education plans required by law.

[2] The *inclusiveness principle* states that NAEP should, to the degree technically, ethically, and financially possible, assess an inclusive sample of all children in the designated age or grade populations. The *quality principle* addresses the necessity for NAEP to be exemplary in the development and use of assessment and reporting techniques and practices that produce reliable, fair, and valid results.

Questionnaire administered for the first time in 1994 and then describes the research questions, methods, and findings of the Panel's IEP study. A parallel section on LEP students is placed next, and the chapter concludes with a brief discussion of the procedural changes for IEP and LEP students that have been introduced by NAEP since 1994.

## *Background for Panel Studies*

### *Exclusion Procedures in Effect through 1994*

As noted in chapter 3, TSA samples of students within schools are drawn initially by requesting the participating schools to provide comprehensive lists of all students at the target grade level. Students from this list are then randomly selected by the NAEP contractor, and the names of selected students are sent back to the schools. At this point, local personnel are asked to identify any IEP or LEP students drawn in the sample and to make exclusion decisions regarding these children. The decisions are made based on written exclusion guidelines provided by NAEP.

The guidelines in effect from 1990 through 1994 (shown in figure 4.1) stipulate that IEP students may be excluded if 1) they are mainstreamed less than 50 percent of the time and are judged incapable of participating meaningfully in the assessment, or 2) the IEP team or equivalent group has determined that the student is incapable of participating meaningfully in the assessment.[3] With regard to LEP students, the criteria state that students may be excluded if they are native speakers of a language other than English, have been enrolled in an English-speaking school for less than two years, and are judged to be incapable of taking part in the assessment. If the student is in a bilingual education program, the two-year time limit is waived. Finally, in reference to both groups of students, the criteria instruct that the students should be assessed if, in the judgment of school staff, they are capable of taking the assessment. In cases of doubt, the school should err on the side of inclusion.

[3] Mainstreaming is generally interpreted to mean that the student is served in the regular classroom and receives at least some of the same instructional content as is offered to other students. The IEP team comprises educators from the school and district who decide on appropriate education choices for each student with disabilities. These decisions are recorded in the student's individualized education plans.

**Figure 4.1.** ***Criteria for excluding students from NAEP assessments: 1990-1994***

***The intent is to assess all selected students. Therefore, all selected students who are capable of participating in the assessment should be assessed.***

Some of the students identified on the Administration Schedule as Limited English Proficient (LEP) or as having an Individualized Education Plan (IEP) may be incapable of participating meaningfully in the assessment. The assessment administrator, with the advice of staff members knowledgeable about the IEP/LEP students, may exclude such students, as described below.

1. A student identified on the Administration Schedule as LEP may be excluded from the assessment if he/she:

   - ◆ Is a native speaker of a language other than English;

   AND

   - ◆ Has been enrolled in an English-speaking school (not including a bilingual education program) for less than two years;

   AND

   - ◆ Is judged to be incapable of taking part in the assessment.

2. A student identified on the Administration Schedule as having an IEP or equivalent classification may be excluded from the assessment if:

   - ◆ The student is mainstreamed less than 50 percent of the time in academic subjects and is judged incapable of participating in the assessment;

   OR

   - ◆ The IEP team or equivalent group has determined that the student is incapable of participating meaningfully in the assessment.

3. IEP/LEP students meeting the above criteria should be assessed if, in the judgment of school staff, they are capable of taking the assessment.

WHEN THERE IS DOUBT, INCLUDE THE STUDENT.

## *Exclusion Rates*

As can be seen in table 4.1, identification and exclusion rates for IEP and LEP students have remained relatively constant across each of the TSA assessments. In 1990 and 1992, about 9 percent of students were identified as IEP and approximately 5 percent

were excluded. These rates rose somewhat in the 1994 TSA, when 11 percent were identified and 6 percent excluded in the fourth-grade reading trial.[4] With regard to LEP students, the rates have been constant across all three trials, but vary by grade. At the eighth grade, 3 percent to 4 percent have been identified and 2 percent excluded; at the fourth grade, 6 percent have been identified and 3 percent excluded across all of the TSAs.

**Table 4.1.** ***Identification and exclusion rates for IEP and LEP students in the TSAs***

| | IEP | | LEP | |
|---|---|---|---|---|
| | Percent identified | Percent excluded | Percent identified | Percent excluded |
| 1990 Grade-8 mathematics | 9 | 5 | 3 | 2 |
| 1992 Grade-8 mathematics | 9 | 5 | 4 | 2 |
| 1992 Grade-4 mathematics | 9 | 4 | 6 | 3 |
| 1992 Grade-4 reading | 9 | 5 | 6 | 3 |
| 1994 Grade-4 reading | 11 | 6 | 6 | 3 |

These rates are generally similar to the national NAEP identification and exclusion rates for the same years. However, in each year, there also has been considerable variability in identification and exclusion rates across states. In 1994, state-level IEP identification rates ranged from 8 percent to 18 percent, whereas state-level IEP exclusion rates ranged from 2 percent to 9 percent. LEP identification and exclusion rates were, predictably, even more variable between states: in 1994, state-level LEP identification rates ranged from none to 23 percent, while state-level LEP exclusion rates ranged from none to 9 percent. These data are shown in tables 4.2 and 4.3 for IEP and LEP students respectively.

[4] The higher 1994 rates may be due to a number of factors, including the tendency for states to classify increasing numbers of learning disabled students at IEP.

**Table 4.2.** ***1994 TSA identification and exclusion rates for IEP students, by state (based on percent of total sample)***

| | High percent identified (13-18%) | Medium percent identified (11-12%) | Low percent identified (8-10%) |
|---|---|---|---|
| High percent excluded (7-9%) | FL, MD, ME, SC, TX | WV, WI | |
| Medium percent excluded (5-6%) | CT, DE, NH, NM, NC, MA, TN | AL, AZ, AR, CO, IN, LA, MO, VA, UT | CA, GA, MI, MS, NY, PA |
| Low percent excluded (2-4%) | NE | IA, MT, RI, WA WY | HI, ID, KY, MN ND, NJ |

**Table 4.3.** ***1994 TSA identification and exclusion rates for LEP students, by state (based on percent of total sample)***

| | High percent identified (11-23%) | Medium percent identified (3-6%) | Low percent identified (0-2%) |
|---|---|---|---|
| High percent excluded (5-9%) | CA, TX | | |
| Medium percent excluded (2-3%) | AZ | CO, CT, FL, MA, NJ, NM, NY | |
| Low percent excluded (0-1%) | | HI, ID, RI, WA | AL, AR, DE, GA, IN, IA, KY, LA, ME MD, MI, MN, MS, MO, MT, NE, NH, NC, ND, PA, SC, TN, UT, VA, WV, WI, WY |

Although these large variations in state-level exclusion rates are suggestive, without additional information it is impossible to determine whether the observed variation reflects true population differences or simply differences in the states' special education policies and practices. In the latter case, some states' mean NAEP performance scores could be artificially increased or decreased as a consequence of their different interpretations of the exclusion guidelines.

## *Characteristics of IEP Students Sampled for the 1994 TSA*

In addition to student data, NAEP collects related questionnaire data from teachers and other school personnel. In 1990 and 1992, teachers provided limited questionnaire data on the characteristics of excluded IEP and LEP students. At the Panel's recommendation, this data collection was expanded in 1994 to include all sampled IEP and LEP students, whether or not they participated in the assessment. The new IEP/LEP Questionnaire data provided a general picture of the IEP population sampled by NAEP and allowed for some simple comparisons between included and excluded students. Some of the findings, as they pertain to IEP students, are presented below. Findings for LEP students are shown in a later section of the chapter.

Across all of the states in the 1994 TSA, nearly 129,000 students were initially selected to participate. Eleven percent of these were identified as IEP due to disability,[5] including 6 percent who took the assessment and 5 percent who where excluded. Among all these IEP students, approximately 71 percent were classified as learning disabled. Most of the remaining students were divided among the disability categories of mental retardation (7 percent), speech impairment (8 percent), emotional disabilities (6 percent), and multiple disabilities (3 percent); orthopedic impairments and sensory handicaps accounted for 2 percent, and the remainder were classified as "other." [6]

By teacher report, the modal reading level was third grade for the IEP students who took the assessment and second grade for the IEP students who were excluded; there was considerable overlap between included and excluded students on this dimension however. Finally, the percent of time mainstreamed was 50 percent or more for nearly all of the included IEP students. Contrary to what one might expect from the exclusion criteria however, only 29 percent of the *excluded* students were reported to spend less than 50 percent of their school day in a regular class (i.e., mainstream) setting. This may reflect recent trends toward serving special education students in the least restrictive educational setting, it may suggest a level of noncompliance with the exclusion criteria, or it may simply highlight the effects of using different wordings in the exclusion criteria and in the questionnaire. (The criteria speak of percent time mainstreamed in academic subjects and the questionnaire only asks about the percent of the school day spent in a mainstream setting.)

IEP/LEP Questionnaire data collected by NAEP in 1994 represented an improvement over data available in the past. However, they still left many questions unanswered. In particular, they provided no objective measure of reading ability on which to evaluate excluded students, and they did not directly address questions about the implementation of the exclusion process or the kinds of adaptations and accommodations that might be used to increase inclusion for this population.

[5] In some jurisdictions, gifted students also receive IEPs. All of the statistics presented here refer to students whose IEP is based upon disability.

[6] These and other results reported from the IEP/LEP Questionnaires are representative of all students with disabilities in the regular school fourth-grade populations of the participating states. They do not, however, include students in special education schools because the latter are excluded from the NAEP samples. Thus, some proportion of the most seriously impaired students are not represented in the estimates.

## *The Panel's Study of Assessability and Exclusions among IEP Students*

The Panel's own study of students with disabilities was designed to address four primary research questions:

1. What is the assessability of the currently excluded IEP students on the NAEP reading assessment?

2. What accommodations, if any, would be needed to include additional IEP students?

3. How was the exclusion decision process for IEP students implemented in the 1994 TSA?

4. Was the IEP exclusion process implemented in a comparable manner in different states? More specifically, were there reading levels at which a student was likely to be included in some states but excluded in others?[7]

The study questions were addressed by collecting independent information on a sample of students selected for the 1994 TSA who also had been identified by their local schools as having individualized education plans. Some of these students had actually participated in the TSA, while others had been excluded on the basis of their disability. For the Panel's study, site visitors met with the students to obtain measures of reading achievement; they also interviewed the students' teachers and local NAEP assessment administrators to gather additional information about the students and, more generally, about the implementation of the exclusion process. The study was conducted with 123 schools in four states and involved 416 students.

### *Assessability*

The reading achievement of the sampled students was measured using the Woodcock Johnson Broad Reading Cluster (WJ BRC), a brief adaptive test that is well regarded by special education researchers and accommodates a wide range of student capabilities.[8] WJ BRC results were then compared to the estimated NAEP scores (plausible values) for those students in the sample who had also taken the 1994 NAEP reading assessment. In particular, the intent was to determine whether the relationship between the two reading measures changed for different levels of reading

[7] F. Stancavage, D. McLaughlin, R. Vergun, C. Godlewski, and J. Allen, "Study of Exclusion and Assessability of Students with Disabilities in the 1994 Trial State Assessment of the National Assessment of Educational Progress," in *Quality and Utility: The 1994 Trial State Assessment in Reading, Background Studies* (Stanford, CA: The National Academy of Education, forthcoming).

[8] R.W. Woodcock and M.B. Johnson, *WJ-R Tests of Achievement: Standard and Supplemental Batteries* (Chicago: The Riverside Publishing Company, 1989). See also Gary L. Hessler's *Use and Interpretation of the WJ-R* (Chicago: The Riverside Publishing Company, 1993).

achievement (as measured by the WJ BRC) and, more specifically, whether there was a WJ BRC score level below which the NAEP no longer captured any of the observed variation in the achievement distribution.

The final results suggest that meaningful performance on NAEP is possible for fourth-grade students with grade-equivalent WJ BRC scores at or above 2.1. Because this cutpoint falls very close to the bottom of the achievement range for which comparative data were available (i.e., very near the bottom of the distribution for IEP students who were included the 1994 TSA), it is difficult to confirm the 2.1 cutscore precisely. Nevertheless, *the available evidence is sufficient for the Panel to conclude that the large majority of fourth-grade students with disabilities are assessable on the current NAEP instrument if the goal is to achieve a level of measurement that would allow information about these students to contribute to estimates of states' overall performance.*

In the study sample, which was representative of students with disabilities (attending regular schools) in four TSA states, it appeared that 83 percent were assessable, including 93 percent of those who had been assessed in 1994, and 70 percent of those who had been excluded. The Panel's estimate, although derived on an entirely different basis, is remarkably similar to the estimate of 85 percent assessable reported by the National Center on Educational Outcome (NCEO).[9]

*The study data also indicate, however, that the current NAEP reading test is not particularly well suited to the reading abilities of many IEP students, in particular, those that are reading a grade or more below grade level.* A more appropriate measure for these students would address the same reading outcomes but be based on less difficult reading passages.

Three recommendations flow from these conclusions.

1. NCES and NAGB should continue efforts to encourage greater participation of students with disabilities in the current NAEP assessments.

2. Results for students with disabilities assessed under standard conditions should be aggregated with results for all other students in producing the overall and subgroup achievement estimates normally reported for the nation and the states. However, results for the population of students with disabilities should *not* be disaggregated or reported separately. Although good information on the achievement of these students would certainly be useful to those interested in the educational progress of students with disabilities, the Panel is concerned that the disaggregated data would not be sufficiently precise to prove helpful, and might even be misleading. Furthermore, because the aggregate population of students with disabilities changes over time in response to changes in special education policy or funding, it would be extremely difficult to track meaningful performance trends for this category of students as a

[9] National Center on Educational Outcome, *Synthesis Report 15: Recommendations for Making Decisions about the Participation of Students with Disabilities in Statewide Assessment Programs* (Minneapolis, MN: Author, 1994), 5, and private communication, Kevin McGrew, NCEO Senior Researcher, May 23, 1995.

whole. For example, as was pointed out earlier in this chapter, identification rates for IEP students in the TSA samples increased from 9 percent to 11 percent between 1992 and 1994.

3. As recommended in chapter 2, NAEP should also work to develop assessments that can measure accurately over a broader range of student proficiency levels and thereby provide better estimates at both ends of the achievement distribution. For efficiency, such an assessment would almost certainly require some adaptive mechanism for matching students with assessment tasks appropriate to their level of proficiency.

## *Accommodations*

One of the topics addressed at length in the teacher interviews concerned the kinds of accommodations that these students received in their schools and the kinds of accommodations that might be needed to allow them to participate meaningfully in NAEP. Responses indicated that the teachers were inclined to be quite liberal in recommending accommodations (or exclusions) for students with disabilities. This may be due to the fact that teachers are most used to making decisions for instructional settings, in which case the goal would be to facilitate the student's learning in every possible manner. They are less used to weighing the needs for standardization in large-scale assessments. In any event, the teachers recommended assessment accommodations for 53 percent of the students in the study sample, including 43 percent of excluded students and 62 percent of those who had participated in the 1994 TSA. They also recommended excluding a small percentage of those who had participated, and slightly more than half of those who had not. *Had the teachers' recommendations been followed, only 20 percent of the 416 IEP students in the Panel's study would have been assessed without accommodations in the 1994 TSA, as opposed to the 56 percent who actually were assessed.* This finding is consistent with the experience of the 1995 NAEP field test, which also found that accommodations were provided for a significant portion of the students who otherwise would have been included without accommodation.

Among the learning disabled students who accounted for 67 percent of the Panel's study sample, the most commonly suggested accommodations were

- Shorter tests and/or more time for testing (85 percent);
- Oral reading of directions with or without interpretation (74 percent);
- Testing in small groups or special education classes (70 percent); and
- Oral response with or without assistance and interpretation (69 percent).

There was a great deal of diversity in the specific accommodation recommended—for example, shortening the test versus extending the testing time versus allowing beaks in testing. The majority of the recommended accommodations, however, appeared to fall into a few broad categories, and many could prove to be interchangeable in their effects. Standardizing accommodations—to the extent feasible while still meeting

individual student needs—would facilitate study of the impact of accommodations on performance and hence, ultimately, advance the goal of scaling and reporting NAEP results for accommodated students together with the main population.

The Panel's study was not designed to address the impact of accommodations on test performance systematically, and the results provide only limited insight into the problem. During the site visits conducted for the study, students who had grade-level scores of at least 2.6 on the WJ BRC read a NAEP reading passage taken from a released 1992 reading item block and answered five of the items that had been administered with that block, including three multiple-choice items and two items for which the students wrote their own answers.[10] Among other analyses, the actual difficulty levels of these items, as they had been when the items were administered under regular NAEP conditions in 1992, were compared to the difficulty levels under the "accommodated" conditions of the Panel's study. (The latter conditions included one-on-one administration, as well as a shortened test.) However, because none of the students in the sample had actually taken the 1992 TSA, and because this particular item block had not been readministered with the 1994 TSA, the statistical comparisons between regular and accommodated conditions could only be approximate. *Nevertheless, the comparisons did not suggest any pronounced, systematic effect of accommodations on achievement results.* The analyses to be undertaken for the 1996 NAEP (in which accommodations were provided for a subsample of IEP students who took the main national assessment) should provide a much better estimate of the impact of accommodations.

For the time being, the Panel recommends that NCES and NAGB continue their present efforts to increase participation of students with disabilities by offering accommodations. However, the Panel also suggests that NCES develop guidelines for the use of these accommodations that, to the extent possible, are aimed at accommodating only those students for whom the standard NAEP administration would be clearly inappropriate.

Furthermore, the Panel suggests that NCES continue research into the impact of accommodations on item performance, and that this research include consideration of the extent to which the same accommodations (e.g., extended time) also improve the performance of nondisabled students. The latter analyses are necessary in order to better determine whether scores earned under accommodated conditions are inflated relative to the scores of other students with which they will be combined. All research into the impact of accommodations will be simplified if, to the extent possible, the choice of accommodations on NAEP can be reduced to the fewest standardized alternatives that adequately address the special needs of these students.

Finally, the Panel notes that there may be certain kinds of accommodations that could confer considerable psychological benefit by making students and teachers more comfortable about participation without compromising the validity and comparability of the results. One such accommodation might be to use staged assessments in which tasks are sequenced in difficulty and stopped when they exceed the student's range of competency. Such a solution could be compatible with adaptive testing, which has already been recommended here and in chapter 2.

[10] The 2.6 cutoff was established prior to the data collection and turned out to be considerably higher than the 2.1 cutoff suggested by subsequent analyses.

## *Exclusion Process*

Another topic addressed in the study interviews with teachers and local assessment administrators was the implementation of the exclusion process. The results indicated that the local school personnel who were responsible for the exclusion process generally followed (to the best of their ability) the written guidelines provided by NAEP. *The single factor most frequently cited as influencing the inclusion/exclusion decision was the child's reading level.* Choosing from a precoded list of options, decision makers for 31 percent of the children cited "reads/doesn't read well enough to take the NAEP" as the primary influence on their decision; an additional 16 percent choose the slightly different option, "reading/not reading at grade level."[11] *Furthermore, percent time mainstreamed and estimates of the student's functional grade level were the only significant predictors of exclusion status in a multivariate logistic regression that included a number of other factors hypothesized to influence exclusion.* The latter included such factors as student race/ethnicity, student gender, student discipline problems, the instructions regarding assessment in the student's IEP, and the types of professionals participating in the exclusion decision process. The greatest barrier to more appropriate exclusion decisions appears to be decision makers' interpretations of the reading achievement level required to participate meaningfully in the NAEP assessment.

The Panel recommends that NCES explore the feasibility of exclusion guidelines based on a prespecified method for estimating whether a student's functional reading level reaches that required for meaningful participation. Such methods might range anywhere from simply embedding sample assessment tasks into the exclusion guidelines package on the one hand, to providing an actual pretest for potentially excluded students on the other. Like some other recommendations already discussed, this one would fit nicely with an adaptive testing model.

## *Comparability Between States*

The Panel's study involved students from four states, selected to maximize the between-state differences on two dimensions: 1) the percentage of students identified as IEP in the 1992 TSA and 2) the percentage of identified IEP students excluded in that year. States were first divided into thirds based on their identification rates, then separately divided into thirds based on the proportions of identified students who were excluded, and finally cross classified on these two dimensions. One study state was selected from each of the extreme cells of the cross classification except for the high-identified-by-high-excluded cell. Because this cell included only one state, Oklahoma, which was not participating in the 1994 TSA, the fourth study state was selected from the adjacent medium-identified-by-high-excluded cell.

Based on data from the standardized WJ BRC reading assessment administered to all students in the study sample, an analysis of between-state differences in exclusion

[11] The NAEP guidelines actually state that the child should be excluded if he or she is determined to be "incapable of participating meaningfully in the assessment;" a fair proportion of respondents apparently operationalized this as "not reading at grade level."

outcomes was undertaken. For this study, the exclusion cutpoint was defined as the reading grade level at which the probability of exclusion reached 50 percent. *Using this definition, states selected for the study were found to differ significantly in their criterion for exclusion decisions. State 1 had the highest cutpoint, with an estimated reading grade level of 3.56. State 2 had the lowest cutpoint, with an estimated reading grade level of 0.85. That means that a substantial number of students who would have been included by State 2 would not have been included by State 1.* (See table 4.3.) The implied exclusion cutpoints for States 3 and 4, which were 2.95 and 3.35 respectively, did not significantly differ from one another.

**Table 4.3.** ***Implied cutscore for exclusion and actual percent of students excluded because of IEP status in the four study states***

| | Implied cutscore for exclusion (WJ BRC grade level) | Actual percent of students excluded |
|---|---|---|
| State 1 | 3.56 | 6 |
| State 2 | 0.85 | 4 |
| State 3 | 2.95 | 6 |
| State 4 | 3.35 | 5 |

In a related analysis, the study used data from the NAEP IEP/LEP Questionnaires to roughly predict reading grade levels and then estimated NAEP scores. Adding the estimated NAEP scores for all of the excluded IEP students who would have passed the approximate cut level for assessability (those with estimated reading levels of grade two or higher) caused state mean scores to decline one-and-a-half to three scale points, suggesting the size of the impact had these students actually been included in the TSA assessment. (See table 4.4.) Predictably, State 2, which appeared to use the most stringent criterion for exclusion, would have declined the least.

**Table 4.4.** ***Estimated impact on mean state scores (expressed as average plausible values) if all students with WJ BRC grade-level reading estimates of 2.0 or higher had been included in the 1994 TSA***

| | Average plausible values for the TSA as actually administered | Estimated average plausible values if all students with grade-level reading estimates of ≥ 2.0 were included |
|---|---|---|
| | (n = 11,071) | (n = 11,550) |
| State 1 | 204.5 | 202.4 |
| State 2 | 221.9 | 220.4 |
| State 3 | 205.2 | 202.1 |
| State 4 | 217.6 | 215.3 |

These findings add force to the Panel's recommendation, stated above, that exclusion guidelines be made more stringent and also be based on less subjective criteria for exclusion.

## *Characteristics of LEP Students Sampled for the TSA*

Approximately 6 percent of the students selected for the 1994 TSA were classified as LEP. Slightly more than 70 percent of these students were Spanish speakers, and about one-quarter attended schools in which more than 60 percent of the students spoke the same non-English language as themselves. As with the IEP population, teachers most frequently reported grade three as the reading grade level for LEP students who had participated in the NAEP assessment and grade two as the reading grade level for those who had not.

About half of the identified LEP students were included in the 1994 TSA reading assessment; the remainder were excluded. Unfortunately, the IEP/LEP Questionnaire did not include a question about number of years in an English-speaking school, so the data offer little evidence with regard to whether the exclusion guidelines were applied appropriately to those LEP students not in bilingual programs. (As noted above, if a student was in a bilingual program, the two-year limit on potential eligibility for exclusion was waived.)

With regard to the prevalence of bilingual programming, the questionnaire data indicate that three-quarters of LEP students spent at least some part of their school day in a special language program and that, among the latter, about one-third (approximately 30 percent of the full LEP population) received reading and writing instruction and/or one or more content courses, such as mathematics, taught in their native language. Those who participated in the native language content courses were unlikely to have taken the TSA (about 70 percent were excluded), but inclusion was more common among the students who had only reading and writing instruction in their native language or received some other (i.e., not bilingual) special language program.

## *The Panel's Study of Assessability and Exclusions among LEP Students*

The Panel conducted a second study of assessability and exclusions among LEP students that had goals and methods similar to its study of IEP students.[12] However, the Panel was unable to identify a standardized measure of English-language reading for the LEP students on which experts on second language learners could agree; thus

[12] F. Stancavage, J. Allen, and C. Godlewski, "Study of Exclusion and Assessability of Students with Limited English Proficiency in the 1994 Trial State Assessment of the National Assessment of Educational Progress," in *Quality and Utility: The 1994 Trial State Assessment in Reading, Background Studies* (Stanford, CA: The National Academy of Education, forthcoming).

no such measure was used. Furthermore, at the recommendation of the Office of Bilingual Education and Minority Languages Affairs (OBEMLA), data were collected only for those LEP students who were in at least their second year of instruction at an English-speaking school.

The research questions for the LEP study were correspondingly modified to reflect the differences in available data. The applicable research questions were.

1. What is the assessability of the currently excluded LEP students on the NAEP reading assessment (limited test)?

2. What accommodations or adaptations, if any, would be needed to include additional LEP students?

3. How was the exclusion decision process for LEP students implemented in the 1994 TSA?

For this study, data were collected from 254 students in three states and 65 schools. The data collection methods generally paralleled those used for the IEP study, and Spanish bilingual site visitors were used throughout. The student assessments were transacted in English only, however, to more closely approximate the conditions of the 1994 TSA.

Neither the states nor the schools sampled for this study were intended to be representative of the nation. Rather, they were selected because they had high proportions of LEP students. Furthermore, at some of the sampled schools, because the numbers of LEP students were in fact very high, site visitors only gathered data for a subset of these students. In these cases, site visitors were instructed to sample equal numbers of included and excluded students.[13] Taken together, these sampling procedures allowed for a sample of adequate size to be gathered most economically, and the resultant data were appropriate for the questions being asked in the study. However, because the sample was purposive rather than representative of the full TSA or even of the particular states or schools that were visited, certain kinds of questions cannot be answered accurately using the study findings. In particular, one cannot use the distributions in the study sample (e.g., percentage in bilingual programs) to generalize to the full population.

Despite these caveats, the sample for the LEP study turned out to be not unlike the full 1994 TSA LEP sample, based on what is known from the IEP/LEP Questionnaire responses discussed earlier. For example, 77 percent of the study sample was Spanish speaking, as compared to 78 percent in the full TSA sample. On the other hand, only 38 percent of the Panel's study sample was excluded from the 1994 TSA, as compared to approximately 45 percent overall exclusions for the TSA.[14]

[13] On average, site visitors were only able to gather data on five or six students per day, and most site visits for both of the Panel's studies were scheduled to be completed in a single day. To accommodate this schedule, sampling of students within schools was necessary in some cases. Only a few of the schools in the IEP study were affected, but sampling occurred in a high percentage of schools in the LEP study.

[14] The difference in percentage excluded could be due to the fact that the study sample was restricted to LEP students who were in at least their second year of instruction at an English-speaking school. The IEP/LEP Questionnaire, however, was completed for all LEP students.

As mentioned previously, the IEP/LEP Questionnaire data provided no information on the years of English-language schooling for sampled students. In the Panel's study sample, 77 percent of the students had attended an English-speaking school for four or more years—essentially their entire school career. Nearly 40 percent of these long-resident LEP students had been excluded from the 1994 TSA: they accounted for 78 percent of the excluded students and 76 percent of the included students in the sample. Relatedly, only 13 percent of the students in the sampled schools were ineligible for the Panel's study because they had been in English-speaking schools for less than two years; this percentage, however, varied across the three states in the study.

## *Assessability*

As noted, no standardized reading measure was used in this study. Rather, to evaluate assessability, the site visitors began with a brief second-grade story that had been developed and tested for oral reading assessment.[15] Students were asked to read the story firstly to themselves, then aloud, and finally, to retell it in their own words. If the students were reasonably successful at the oral reading and retelling task (e.g., were able to read through the story, with or without errors, in a reasonable amount of time), they were then given the same abridged NAEP item block as the students in the IEP study who had scored 2.6 or higher on the WJ BRC. Provisions were made to allow the LEP students to answer the NAEP questions orally if necessary, but this option was never exercised.

In the site visit assessment, 98 percent of the students who had participated in the 1994 TSA and 78 percent of the excluded students did well enough with the second-grade story to progress to the NAEP item block. The passage in the NAEP block is considerably more difficult than the second-grade story that was used for screening. Nevertheless, 86 percent of the included students and 57 percent of the excluded students answered at least one of the five items in the abridged item block correctly. Predictably, fewer of the students were successful with the two questions that required them to write answers in their own words. Thirty-nine percent of the included students, but only 16 percent of the excluded students, received credit for at least one of these items. *Overall, the evidence, although less rigorous than that available for the IEP students, still suggests that a high proportion (probably greater than 75 percent) of these LEP students would have been assessable on the current NAEP instrument.*

The Panel's conclusions regarding assessability are based on the same, relatively undemanding criterion that was used in the IEP study: the estimates of student achievement need only be accurate enough to allow scores for these students to contribute to state averages, not to make conclusive judgments about the achievement of individual students. Furthermore, conclusions about assessability drawn from this definition do not preclude the possibility that achievement, for at least some LEP students, might be underestimated by NAEP because of linguistic demands of the

[15] L. Leslie and J. Caldwell, "What Can I Get for My Toy?" *Qualitative Reading Inventory* (La Porte, IN: Harper Collins Publishers Inc. 1990), 117.

assessment that are not essential to the competencies the assessment is attempting to measure. The latter topic could be better investigated in a different subject area (such as science or mathematics) that is less language dependent than reading.

## *Accommodations and Adaptations*

Teachers of LEP students, like teachers of IEP students, were inclined to be quite liberal in recommending accommodations, adaptations, or exclusions for their students. The teachers recommended assessment accommodations or adaptations for 45 percent of the LEP students in the study sample, including 45 percent of excluded students and 45 percent of those who had participated in the 1994 TSA. They also recommended the exclusion of a small percentage of those who had participated, and slightly more than half of those who had not. *Had the teachers' recommendations been followed, only approximately half as many of the LEP students in our study would have been assessed under standard conditions in the 1994 TSA (30 percent compared to 63 percent).*

The most commonly suggested accommodations or adaptations were

- Extended time and/or shorter versions of the test (82 percent);
- Use of pictures or diagrams in the assessment presentation, with or without instructions in native language, and/or out-of-grade testing (75 percent); and
- Testing alone, in small groups, or in special education classes (52 percent).

In most instances, the teachers recommended that the students respond in English, although some suggested that the responses could be given orally or that the students should be able to receive assistance in interpreting their responses. Perhaps because the subject area of the assessment was reading, there were fewer recommendations for offering instructions in the student's native language (10 percent) or allowing the student to respond orally in that language (4 percent).

## *Exclusion Process*

When queried about their primary reasons for inclusion or exclusion, teachers in the LEP study identified reading level (either "reads/doesn't read well enough" or "reading/not reading at grade level") with regard to 32 percent of the students. Reading level was followed by appropriateness to the child's instructional program (26 percent), whether or not the student was included in the state assessment (17 percent), and ability to understand oral English (12 percent). *The results of a multivariate logistic regression to predict exclusion identified four factors as significant. The first two, the percentage of time per week spent in a special language (bilingual) program and the exclusion from state, district, or other grade-level standardized tests, increased the likelihood of exclusion; the latter two, the number of years spent in a*

*special language program and the teacher's (higher) estimation of the student's functional grade level for writing English, decreased the likelihood of exclusion.*

Perhaps the most disturbing finding from the LEP study was the significant fraction of these LEP students who were excluded from the assessment despite the fact that more than three-quarters had four or more years in English-speaking schools. The Panel acknowledges that NAEP, like other large-scale assessments, may not function in an equivalent manner for students with limited English proficiency or those who are receiving at least some of their instruction in their native language. Indeed, NAEP may not offer these students an optimal opportunity to demonstrate their competence. Nevertheless, these students may too easily drop from sight if they are excluded from major assessment efforts and they may, ultimately, receive less attention to their education needs.

The Panel therefore recommends that NCES continue its efforts to identify appropriate adaptations or accommodations and permit the inclusion of larger proportions of LEP students in the assessments.

## *Changes in Exclusion Policies for the 1996 Assessment*

For 1996, NCES undertook to expand inclusion of IEP and LEP students by modifying the exclusion guidelines and by providing assessment accommodations and a bilingual Spanish version of the assessment for some students in the national sample. Because NAEP has thus far had little experience with analyzing or scaling results obtained under these conditions, and because even minor changes in procedures can affect the interpretation of trends over time, the 1996 samples were subdivided to allow the impact of these changes to be investigated systematically. With regard to the national assessment in mathematics, one-third of the participating schools used the 1994 exclusion criteria and two-thirds used the new criteria. Schools using the new criteria were further subdivided by offering accommodations and bilingual assessments in some and not in others. The design for the national science assessment was similar, except that the old exclusion guidelines were not used for any portion of the sample because there was no trend line to maintain. (A new science framework was introduced in 1996.) Furthermore, Spanish-speaking students were provided with an English-Spanish glossary of scientific terms but not a bilingual version of the assessment.

The design for the *state* mathematics and science assessments did not include accommodations or bilingual assessments for any of the students because such arrangements were not covered in the 1996 participation agreement signed by the states. Moreover, NCES and ETS appropriately wished to work out the practical and analytical requirements for accommodated assessment before scaling up to the state assessments.

The new exclusion guidelines, which necessarily differ somewhat depending on whether or not accommodations and bilingual assessments are provided, are shown in figures 4.2 and 4.3. The greatest change is the emphasis on inclusion rather than exclusion in the phrasing. Several changes have also been made to clarify the guidelines, such as referring to "instruction primarily in English" rather than years in

an "English-speaking school." ***The Panel commends these efforts to increase inclusion, and also the fact that the new procedures are being added in a way that will protect trend lines and improve the interpretation of results. The Panel encourages NCES and ETS to carefully examine the results of these experiments for both IEP and LEP students.***

**Figure 4.2.** ***Criteria for including students in the assessment: 1996 Sample 2 (no accommodations provided)***

**The intent is to assess all selected students. However, since NAEP is a timed assessment administered in English to groups of students, it may not be possible to assess some students with disabilities or Limited English Proficiency.**

Students with Limited English Proficiency

A student who is classified as limited English proficient (LEP) and who is a native speaker of a language other than English should be included in the NAEP assessment unless:

- The student has received mathematics, science, and language arts instruction primarily in English for less than three school years, including the current year;

AND

- The student cannot demonstrate his or her knowledge of mathematics or science in English without an accommodation or adaptation.

Students with Disabilities

A student identified on the Administration Schedule as having an Individualized Education Plan (IEP) or equivalent classification should be included in the NAEP assessment unless:

- The IEP team or equivalent group has determined that the student cannot participate in assessments such as NAEP;

OR

- The student's cognitive functioning is so severely impaired that she/he cannot participate;

OR

- The student's IEP requires that the student be tested with an accommodation or adaptation, and the student cannot demonstrate his/her knowledge of mathematics or science without that accommodation or adaptation.

WHENEVER THERE IS DOUBT, INCLUDE THE STUDENT

Figure 4.3. *Criteria for including students in the assessment: 1996 Sample 3, national*

**NAEP intends to assess all selected students and to provide the most common accommodations or adaptations required for testing students with disabilities or Limited English Proficiency. In some cases, however, it may not be possible to assess some of these students.**

Students with Limited English Proficiency

A student who is classified as limited English proficient (LEP) and who is a native speaker of a language other than English should be included in the NAEP assessment unless:

- The student has received mathematics, science, and language arts instruction primarily in English for less than three school years, including the current year;

AND

- The student cannot demonstrate his or her knowledge of mathematics or science even with the accommodations and adaptations available from NAEP and the school.

Students with Disabilities

A student identified on the Administration Schedule as having an Individualized Education Plan (IEP) or equivalent classification should be included in the NAEP assessment unless:

- The IEP team or equivalent group has determined that the student cannot participate in assessments such as NAEP;

OR

- The student's cognitive functioning is so severely impaired that she/he cannot participate;

OR

- The student's IEP requires that the student be tested with an accommodation or adaptation which NAEP and the school do not provide, and the student cannot demonstrate his/her knowledge of mathematics or science without that accommodation or adaptation.

WHENEVER THERE IS DOUBT, INCLUDE THE STUDENT

Students with disabilities or limited English proficiency account for a significant fraction of U.S. students. In 1994, 11 percent of the students sampled for the TSA in fourth-grade reading were identified as having an individualized education plan related to disability, and 6 percent were identified as limited English proficient. About half the students in each of these groups were judged by local school personnel to be incapable of meaningful participation and were therefore excluded from the assessment.

As laid out in its guiding principles in chapter 1, the Panel believes that the *quality* and *utility* of NAEP rest, in part, on NAEP's broad *inclusiveness,* and that inclusiveness should be encouraged to the degree technically, ethically, and financially possible. For the 1994 evaluation therefore, the Panel undertook a pair of studies to gather information on how additional students might be added to the assessment and how exclusion decisions were currently being made by the schools.

The first of these studies collected new information on 416 IEP students, in four states, who had been sampled for the 1994 TSA. The Panel concluded that more than 83 percent of fourth-grade IEP students were assessable on the current NAEP reading instrument, including 70 percent of those who had been excluded in 1994. Although teachers appeared to be using the correct criteria for their exclusion decisions—the student's reading level and percent time mainstreamed—they tended to assume that participation required a higher level of reading proficiency than was suggested by the Panel's findings. Furthermore, the teachers were very likely to recommend that IEP students be assessed under accommodated conditions. *Their recommendations, if followed, would have increased the total number of IEP students assessed, but considerably lowered the number assessed under standard conditions.* The Panel's study of LEP students, although designed around constraints that somewhat limited the types of conclusions that could be drawn, nevertheless yielded generally parallel findings.

Large proportions of both IEP and LEP students consequently appear to be assessable when the goal is to achieve a level of measurement that would allow information about these students to contribute to estimates of states' overall performance. The current NAEP reading text is not, however, particularly well suited to the reading abilities of the many students in each group who are reading considerably below grade level. Rather, for students near the extremes of the proficiency distribution, some form of adaptive testing that adjusted difficulty without altering the essential framework of the assessment would provide more accurate information as well as a less frustrating assessment experience.

In the 1996 mathematics and science assessment, NCES expanded its efforts to increase inclusion by modifying the local administrator guidelines to stress inclusion, and by providing accommodations and adaptations (including a bilingual version of the mathematics assessment) for students who could not otherwise demonstrate their knowledge of these subjects. The Panel commends NCES and NAGB for these efforts and also for their care in using a 1996 design that protects trend lines and maximizes the technical information that will be available on the impact of accommodations.

The latter is critical if results from students who take the assessment under nonstandard conditions are ultimately to be combined with results from all other students in estimating proficiency levels for the nation and the states.

# 5 Scaling and Analysis of the 1994 Reading Assessment

## Introduction

Most of the assessments used for student evaluation, college admissions, job selection, or state or local accountability are all similar in that they are designed to provide simple but accurate scores for individuals. NAEP, however, is an assessment in which relatively small amounts of information from each of many students are combined to create national and state estimates of achievement over broad subject areas. For example, in the 1994 fourth-grade reading assessment, each student completed only about one-fourth of the total tasks, yet information from all of the tasks—spanning many different types of reading activities—was integrated to provide an overall measure of achievement.[1] In addition, NAEP collects information about basic background characteristics and education practices from the assessed students, their teachers, and their schools.

From its beginnings more than 30 years ago, NAEP has been a survey with broad content coverage. Originally however, it only reported results in terms of student performance on individual items. In reading for example, NAEP might report that 80 percent of fourth-grade students could identify the main idea in a brief fable. Similarly, for mathematics, it might report that 50 percent of eighth-grade students could correctly compute the area of a square when given the area of a triangle contained inside it. These results, when presented alongside the actual text of the items, were concrete and easy to understand. However, there was no systematic way to summarize results from a large number of different exercises in order to derive a general picture of how well U.S. students were doing in reading or mathematics. NAEP consequently began to supplement these item-level results with information that was combined and reported on an overall scale. The essential components of the current scaling design, including the various statistical procedures that support it, were adopted in the mid-1980s when the NAEP contract moved to ETS.

ETS has played a major role in the development and refinement of procedures for NAEP, but NCES has also played a key role. In addition to providing the funding and overall direction for the program, NCES, in its role as the primary federal agency for education statistics, provides guidance in the appropriate use of statistical analysis methodologies and the proper use of data from their national surveys. Thus, NCES not only has a vested interest in how these processes are conducted, but is intimately involved in all aspects of the scaling and analyses used for NAEP.

In general, the process of going from the students' responses on the assessment items to the summary measures of achievement reported by NAEP requires a complex sequence of activities that can be grouped into three main areas:

[1] The proportion of the total item pool addressed by any individual student is even smaller at the higher grade levels; 8th- and 12th-grade students each completed fewer than 20 percent of the total tasks.

1) **Scoring.** Multiple-choice items are scored by machine, while trained raters read and assign scores to the constructed-response items that the students answered in their own words.

2) **Scaling.** Items are mapped onto an overall achievement scale, and statistical adjustments are made so that results from one year can be compared with those from another.

3) **Analysis.** Achievement estimates are generated for populations of interest, such as the United States as a whole, individual regions, and (for the TSA) participating states. Separate estimates are also prepared by gender, race/ethnicity, and other subgroupings of students. Other analyses evaluate the statistical significance of achievement differences between groups and across time, explore relationships between achievement and education practices, and provide more fine-grained interpretations of the various aspects of achievement.

The current design (which is used for both national and state NAEP) is statistically complex. Although complexity has some negative consequences, such as adding time and expense to the analysis and reporting cycle, these shortcomings of the design have historically traded off successfully against its effectiveness for increasing information while reducing the burden on individual students. In recent years, the balance of costs and benefits may be shifting due to programmatic changes that have compounded the impact of the design's statistical complexities. To cite one example of these changes, the overall measurement burden has greatly increased since the inception of state NAEP because virtually the entire scoring, scaling, and analysis process must now be repeated separately for each state. Another example is the increased demand for innovations in assessment design and technology—these have been introduced at such a rapid rate that analyses do not become routine, but change in a variety of ways with every assessment cycle.

Some of the pressures for change have come from the content specialists who, since 1990, have introduced a number of innovations in the kinds of tasks presented to students with the aim of increasing the validity and real-world relevance of the assessment. A great deal of work has been needed to accommodate these content changes statistically, particularly to accommodate the new scoring guidelines that allow different degrees of credit depending on the quality of the student response. Shortly, a policy change affecting a different aspect of the assessment will lead to a new round of exploratory analyses, and potentially to modifications to the scaling model, in order to accommodate the increased numbers of students with disabilities or limited English proficiency who participated in the 1996 assessment.

These compounding complexities have raised various concerns including

- Delays in the time to reporting and
- Increased likelihood of error.

The likelihood of error increases if the assumptions of NAEP's complex measurement models are not satisfied, or their consequences not fully understood. Error can also

result from simple mistakes, such as programming errors, if they are not detected through the normal quality control process.

During the process of evaluating the 1994 TSA in reading, two independent events highlighted the error potential of the system. First, the results for the 1994 national reading assessment, which were being produced concurrently with the TSA reading results using very similar models and procedures, exhibited a statistically significant and unanticipated drop in reading achievement at grade 12. Extensive checking and analysis were required to determine whether some or all of this decline could be explained by methodological problems.[2] Second, two separate and unrelated errors were discovered almost simultaneously in the procedures used to calculate the results.[3] Although these errors were relatively minor in their impact, both affected the already-reported 1992 results in reading and mathematics in addition to the 1994 results.

In this chapter, the Panel

- Provides an overview of the analytic procedures used by NAEP;
- Reviews and summarizes major innovations in the NAEP scaling and analysis approach;
- Describes the problems that occurred in 1994; and
- Offers recommendations for improving the scoring, scaling, and analysis process in the future.

Throughout the chapter, issues are raised and suggestions made with a focus on the NAE Panel's *quality principle:* "NAEP should be exemplary in the development and use of assessment techniques and practices that produce reliable, fair, and accurate results."

## *Overview of NAEP Scoring, Scaling, and Analysis Procedures*

Broadly speaking, the scoring, scaling, and analysis procedures for the 1994 TSA in reading were the same as those that had been used for previous TSA assessments. These procedures encompass a wide range of very specialized tasks that employ a large number of contractor staff and must be conducted in a prescribed sequence over

[2] This event is summarized in L.V. Hedges and R.L. Venesky, "The 1994 Reading Anomaly: Report to The National Academy of Education on the Drop in NAEP Main Assessment (Short-Term Trend) Scores" in *Quality and Utility: The 1994 Trial State Assessment in Reading, Background Studies* (Stanford, CA: The National Academy of Education, forthcoming).

[3] The first error was related to how omitted responses were treated in a computer program used to scale the NAEP results. The second error involved an incorrect mapping of the achievement levels onto the NAEP reading scale. Further description of both errors can be found below and in P.L. Williams, C.M. Reese, J.R. Campbell, J. Mazzeo, and G.W. Phillips, *NAEP 1994 Reading: A First Look,* revised ed. (Washington, DC: National Center for Education Statistics, October 1995), 61.

a period of months. All of the activities lead up to one ultimate outcome: to provide the results that will be reported from the assessment.

The first step involves scoring the students' written responses. This is an extremely labor intensive process that must be completed before the scaling and analyses can begin. Furthermore, during this time, the questionnaire data obtained from students, teachers, and school administrators must be compiled and edited. One of the complexities of the NAEP design arises from the fact that there are important interdependencies among the different types of data (e.g., student assessment responses, teacher questionnaire responses, and so on). As a result, none of the core analyses can be performed until all the data have been scored, entered onto data files, and checked for errors.

The subsequent scaling and analysis steps involve verifying the statistical soundness of each of the assessment items, relating the items to the overall reading scale, establishing the links between years that allow 1994 reading results, for example, to be compared to those from 1992 and, finally, preparing achievement estimates for each of the populations and groups reported. In general, for any given subject area, scoring, scaling, and analysis activities for both the state and national assessments must be conducted simultaneously and cross checked against one another.

## *Scoring*

For the most part, the contractor employs standard procedures for scoring constructed-response items typical of those implemented for other large-scale assessments. As discussed in chapter 2, initial plans for scoring are made while the items are being developed. At that time, preliminary scoring guides are written to accompany each item. Once actual student responses are received, the scoring guides are refined and elaborated by expert raters, and examples of student responses for each score category are inserted.[4] Large numbers of readers are hired to work intensively over a period of several weeks and provided with extensive, standardized training on each item that is to be scored. Throughout the scoring period, continuous checks are carried out to ensure that the responses are being scored in a consistent manner, and a number of statistical indices are calculated that also monitor score reliability levels.

During the 1990s, NAEP has steadily increased the use of constructed-response items in its assessments and this, in turn, has increased demands to score higher volumes of student responses without slowing down the analysis and reporting schedule. At the same time, the mix of constructed-response items has also been shifting to include more items that require extended responses from students. The latter are scored using a graded response scale that gives more or less credit to responses, depending on how well the responses satisfy the requirements of the assessment task.

[4] See appendix A for examples of two scoring guides; one for a dichotomously-scored (right/wrong) item and one for an extended-response item with four score levels.

The scaling process takes the information collected from all of the items administered and summarizes this information onto the subscales defined in the framework. The separate subscale scores are then combined in a weighted formula to reflect their relative importance in the total set of competencies expected of students at each grade level, overall proficiency scores are computed, and achievement estimates are produced for each of the groups and subgroups that will be reported. This process is carried out separately for the TSA and for the national assessment. The scales are then linked together so that state results can be compared to national. In addition, whenever short-term trend data are being reported (as they were for reading in 1994), the scales for the two trend years must also be linked.[5]

Before the actual scaling begins, all items are checked to confirm that they are statistically sound and that they fairly measure achievement for students from different backgrounds. The relationship of each item to the scale is then established and, if the item is to be used to link performance across two assessment years, analyses are conducted to confirm that the relationship of the item to the scale remained the same across the two years.

Within the sampling design used by NAEP, students do not answer enough questions about any specific topic to provide reliable information about individual performance. However, the many analyses that must be completed for the initial reporting, in addition to those that may be carried out later by independent researchers working with NAEP data, are more straightforward if individual "scores" can be estimated and used in the analyses. To do this, information from the achievement items completed by a given student is combined with background information for the same student—taken from the student, teacher, and school questionnaires—to produce estimates of the student's "score." In this way, information about group achievement (which is more accurate because it is based on a much larger set of achievement items) is distributed across the individuals who make up the group (i.e., is distributed across individuals who are characterized by similar responses on the student, teacher and school questionnaires). The process by which this is accomplished in NAEP is commonly referred to as "conditioning," or "conditioning on the background variables," and is conducted separately for each state in the TSA, again adding to the length and complexity of the overall analyses.

Finally, a distribution of possible values reflecting the uncertainty introduced by measurement error is estimated for each student. Five random draws ("plausible values") are taken from this distribution and are retained for further analyses. In this way, the analyses account for the fact that the individual "scores" are imperfect estimates of any given student's actual achievement. The amount of variation among the plausible value estimates differs for each student and is a measure of the uncertainty in that student's estimated achievement. For example, achievement for

[5] All cross-year linkages are constructed using national, not state, NAEP data. In 1994, for example, 1994 national reading results were first linked to the national reading results for 1992. Then the TSA results for 1992 and 1994 were each made comparable to the fourth-grade national reading results for the same year. After that, 1992 and 1994 TSA results could be compared directly.

students who are "average" in their abilities can be estimated more precisely than achievement for students whose performance places them at the very top or very bottom relative to their peers.

### *Analysis*

After scaling is completed and the plausible values have been produced, further analyses are undertaken to produce the tables and comparisons that appear in the NAEP reports. These include producing estimates of achievement for the various subgroups of interest, calculating the statistical significance of differences between the groups (e.g., determining whether the average reading proficiency estimated for students in Iowa is reliably greater than the average reading proficiency estimated for students in West Virginia), and exploring the relationships between the separate dimensions of proficiency or between proficiency and the various characteristics of the students and their education experiences.[6]

## *Recent Innovations in NAEP Methodologies*

Since the last major design change for NAEP, which occurred when ETS won the contract in 1983, NAEP has used an analysis and scaling approach that employs Item Response Theory (IRT) as its primary methodology. Although IRT has successfully and consistently been used over the years, the specifics of the implementation have been constantly refined. A brief summary of the major innovations that have been made to the NAEP scoring, analysis, and scaling procedures, beginning with the 1990 TSA, is presented below.

1) **Technical and procedural activities related to state NAEP.** For 1990, the year of the first TSA, NAEP analysis procedures had to be adapted to handle a much larger amount of data overall and to carry out separate analyses for each of the participating states and jurisdictions in addition to the national sample. Additionally, procedures had to be developed for linking the state results to the national results and for calculating the statistical significance of observed score differences between pairs of states. In the latter case, standard significance tests were adjusted to reflect the fact that, when so many different comparisons are being made, some will be statistically significant by chance alone. The same reasoning applies when one tests for significant gains (or losses) in the same state across time, because the cross-time comparisons are being done simultaneously for all participating states. The particular adjustment adopted by NAEP for both cross-state and cross-time comparisons is called the Bonferroni adjustment.

[6] More detailed descriptions of NAEP scaling and analysis procedures can be found in the *Technical Report of the NAEP 1994 Trial State Assessment Program in Reading*, December 1995, and *The NAEP 1994 Technical Report*, forthcoming. In addition, a number of informative articles about NAEP measurement issues have appeared in the American Educational Research Association's *Journal of Educational Statistics* (see 17 [2] [Summer 1992]) and the National Council of Educational Measurement's *Journal of Educational Measurement* (see 29 [2] [Summer 1992]).

2) **New content frameworks.** In the 1990s, NAEP has undertaken an extensive review and revision process of the content for every subject area to be assessed. The 1990 TSA in mathematics was based on a new framework, as was the 1992 TSA in reading and the U.S. history and world geography assessments administered nationally in 1994. The new frameworks included, among other things, an increased use of constructed-response items and more emphasis on the use of reporting subscales that capture different dimensions of overall proficiency. A higher volume of analysis was needed to compute and examine results for each subscale, in addition to results for the overall composite scales, and to provide detailed breakdowns of these results for use in the state and national reports.

3) **New scaling methods and models.** New models also were required to scale the variety of item types introduced by the new frameworks and to combine all of these onto a single scale. Beginning in 1992, three separate scaling models were used: one for traditional multiple-choice items, one for dichotomously-scored (right/wrong) constructed-response items, and one for the longer constructed-response items that, as noted earlier, receive different scores depending on the quality of the student's answer. The last of these—referred to as the partial-credit IRT model—required substantial new development on the part of the NAEP contractor.[7]

4) **Introduction of achievement levels.** Beginning with the 1992 assessments, achievement levels were introduced as the primary means for reporting.[8] The achievement levels were developed by expert judges working under NAGB and their contractor, American College Testing (ACT), with statistical input from ETS concerning the items. First, the relationship of the items to the overall proficiency scale had to be quickly determined by ETS and communicated to ACT. ACT then used this information in standard-setting meetings to convert the judges' ratings of individual items into cutpoints on the NAEP scale. Next, these cutscores between achievement levels had to be incorporated into the analysis programs used by ETS, and results by achievement level (percentages of students achieving at or above each cutscore) had to be computed and reported for every jurisdiction and each of the main NAEP reporting groups. These steps required a great deal of communication and coordination between the contractors, and added much more work to be done in a short amount of time.[9]

[7] Multiple-choice and dichotomously-scored constructed-response items are both scored as "right" or "wrong," and are scaled using models that consider how the likelihood of answering the item correctly changes as the student's overall proficiency increases. Separate scaling models are required for multiple-choice items, which a student may sometimes answer correctly by means of random guessing, and constructed-response items, wherein guessing is not possible. For partial-credit items, the scaling model estimates the relationship between score and overall proficiency separately for each of the item's score categories.

[8] Achievement levels had also been implemented in conjunction with the 1990 TSA, but they were viewed as a trial effort and were reported separately.

[9] Here we have focused on the interaction between the achievement levels and the main NAEP analyses. Achievement levels and the level-setting process are discussed further in chapter 6.

5) **Analysis and reporting for short-term trends.** In 1992, NAEP also began reporting national and state results that included short-term trends. For the first time, procedures were implemented to allow results from the main cross-sectional assessment from one year (1992) to be compared to those from another (1990). These procedures, which also were used to link the 1994 reading results to those from 1992, were more complicated than the procedures NAEP had been using (and continues to use) to calculate long-term trends. The latter are based on content frameworks, item types, and administration procedures that have been essentially unchanged for many years. By contrast, the new short-term trends are based on assessments that are being enhanced, in many small ways, with every administration.

6) **Expansion of partial-credit scoring to include short constructed-response items.** For the 1994 reading assessment, NAEP began using a three-point score scale for many of the short answer constructed-response items that previously had only been scored dichotomously. This innovation, which increased the amount of information obtained from students' responses, also expanded reliance on the partial-credit IRT model, discussed above in (3).

7) **Imaging technology.** Starting with the 1994 assessment, NAEP began using imaging technology in its scoring activities. Students' written responses were scanned onto computers and then scored and processed as screen images rather than hard-copy documents. This new technology enabled NAEP to score the constructed responses more efficiently and much more quickly than did the previous paper-and-pencil approach. In particular, the decreased logistic burden allowed NAEP to switch to a procedure in which scorers worked on only a single constructed-response item at a time rather than working back and forth through all of the contructed-response items that appeared together in an item block. By permitting scorers to focus in this way, both consistency with scoring guides and comparability between scorers went up, thus increasing score reliability.

Although the Panel applauds these innovations, two of them are worthy of comment with regard to what the Panel sees as shortcomings. First, as is well known, the Panel has taken a strong position against the use of the current achievement levels for reporting NAEP results. The history of their development and the Panel's arguments against their use is the topic of chapter 6 and is therefore not discussed here. Second, whereas it is important to adjust significance levels when making multiple comparisons, the Bonferroni procedure currently being used by NAEP is not necessarily the most appropriate. As the number of comparisons increases, the power of the Bonferroni adjustment rapidly decreases, meaning that the significance tests may actually be over corrected and the number of statistically significant differences seriously underestimated.

One alternative for controlling the "false positives" introduced by Benjamini and Hochberg[10] has been tested with results from the NAEP TSA.[11] As an example of how results adjusted by the two different methods compare, consider the 34 states that participated in both the 1990 and 1992 eighth-grade mathematics TSA. Using the Bonferroni method, NAEP found and reported significant gains for four states: Rhode Island, Minnesota, Hawaii, and North Carolina. The alternative Benjamini-Hochberg approach showed significant gains for the same four states, but also identified seven others: New Hampshire, Iowa, Colorado, Texas, Idaho, Arizona, and Kentucky.

When reporting NAEP results, the failure to detect and report statistically reliable differences is a serious shortcoming. Therefore, the Panel recommends that NCES consider the adoption of an adjustment for multiple comparisons that is more powerful than the current Bonferroni procedure for detecting significant differences.

*In summary, NAEP has undergone a number of important innovations and changes since 1990, with many beneficial outcomes (e.g., obtaining more complete information on students' performance by adding the use of constructed-response items). The Panel commends NAGB and NCES for these significant improvements. The rate at which such innovations have been introduced may be straining the entire scoring, scaling, and analysis system however, and therefore may be increasing the likelihood of errors or anomalous findings.*

## *Investigation of Decline in Reading Achievement at Grade 12*

While the preliminary results for the 1994 national reading assessment were being reviewed, attention was focused on a statistically significant decline in scores at grade 12. As described in (5) above, trend analyses using the main NAEP assessments were new beginning in 1992 (for mathematics), and 1994 was the first instance in which cross-year comparisons were based on an assessment built to the new reading framework. Consequently, there was no ready comparison against which to gauge the stability of the trend measurement or the real significance of a decline of the observed magnitude. In light of this fact, and the rapid rate of recent innovations and the complexity of the analyses, there was considerable concern that the decline be verified as "real" before the results were released. ETS, of course, carried out a number of checks and analyses to address this question. In addition, NCES felt the need for a second, independent look at the data. NCES therefore asked the NAE Panel to quickly conduct such a review and advise on the accuracy of the results. The

[10] Y. Benjamini and Y. Hochberg, "Controlling the false discovery rate: A practical and powerful approach to multiple testing," *Journal of the Royal Statistical Society* B (55) (1995), 289-300.

[11] V.S.L. Williams, L.V. Jones, and J.W. Tukey, *Controlling error in multiple comparisons with special attention to the National Assessment of Educational Progress: Technical Report No. 33* (Research Triangle Park, NC: National Institute of Statistical Sciences, 1994).

study commissioned by the Panel[12] was carried out and reported to NCES during the weeks preceding the release of the *First Look* reading report[13] in early 1995.

Based on this study, the Panel concluded that it was not possible to determine whether or not methodological problems were a factor in the decline. It was also not possible to be certain whether or not the observed drop in 1994 reading scores was real (i.e., accurately measured). Although no definitive reason or methodological factor related to the decline was found, the intensive review highlighted a number of possible contributing factors, such as normal variations in sampling, under-representation of constructed-response items in the common pool of items used to equate (or link) the 1994 and 1992 assessments, changes in scoring procedures for constructed-response items (see [6] and [7] above), and failure to take account of measurement uncertainty introduced by equating when calculating whether the drop was statistically significant (i.e., greater than would be expected by chance). Although the researchers commissioned by the Panel concluded that none of the possible "culprits" they examined could, by itself, fully account for the score differences, they considered it plausible that several, combined, could have contributed to the magnitude of the overall differences.

For example, the researchers examined the possibility that between-year equating may have been made more difficult as a consequence of changes implemented in 1994 for scoring the constructed-response items. A review of the scoring procedures found that, compared to 1992, the 1994 scoring was more reliable and adhered more closely to the scoring guidelines. As a result, many of the constructed-response items evidenced a different relationship to the overall proficiency scale in 1994 than in 1992 and had to be dropped from the equating set of items. Multiple-choice items were consequently over-represented in the items that remained for equating purposes, compared to the mix of item types for the test as a whole. Because the different types of items may measure different aspects of reading proficiency, this has the potential to reduce the accuracy of the equating.

Furthermore, in the process of doing research on the drop in 12th-grade reading scores, the Panel determined that there is no standard error reported for the equating; as a result, one cannot reliably determine the "goodness" of the equating. Although other studies have investigated equating in complex NAEP-like designs,[14] additional studies are needed to determine the specific effects on NAEP results when equatings are carried out under less than optimal conditions. For these reasons, the Panel recommends that a study of equating error be undertaken to estimate the amount of uncertainty that is introduced into cross-year comparisons under both ideal and less than ideal circumstances (e.g., when it is not possible to reflect accurately the composition of the entire assessment in the set of items used for equating, as happened in 1994). This research should provide a better estimate of the amount of equating error involved, as well as the standard error around that estimate.

[12] L.V. Hedges and R.L. Venesky, op. cit.

[13] P.L. Williams, C.M. Reese, J.R. Campbell, J. Mazzeo, and G.W. Phillips, *NAEP 1994 Reading: A First Look* (Washington, DC: National Center for Education Statistics, April 1995).

[14] For example, see J.R. Donoghue and J. Mazzeo, "Comparing IRT-based Equating Procedures for Trend Measurement in a Complex Test Design," Educational Testing Service. Paper presented at the annual meeting of the National Council on Measurement in Education in San Francisco, CA, April 1992.

One potential basis for evaluating the reality of the observed decline on the 12th-grade short-term trend results would be to look for similar patterns of declining performance in the long-term trend results for the same year. Unfortunately, the 1994 long-term data were not available to the investigators commissioned by the Panel because these analyses had been set aside in favor of other priorities with more urgent reporting deadlines.[15] Furthermore, no systematic efforts had been made to gather data from state assessments or other testing programs that also could have been used to check the reasonableness of the short-term results. Thus, NCES was placed in the uncomfortable position of authorizing the release of the short-term trend results for reading without being able to check them against long-term trends or other related results. The Panel recommends that greater efforts be made to complete related analyses of NAEP data (e.g., short- and long-term trends in the same subject) by the time the national and state NAEP data are released so that results can be cross checked for reasonableness. The Panel further recommends that relevant data sets external to NAEP also be analyzed routinely to help evaluate significant changes in NAEP performance, *whether downward or upward*. In order to avoid further delays in reporting, the data to be used in these various comparisons should be readied in advance, to the greatest extent possible.

Among these alternative sources, long-term NAEP trend results are probably the most comparable because they represent a very similar population. Other indicators that might also be examined include school dropout rates, performance on standardized tests, and trends in college admissions test scores (e.g., Scholastic Assessment Test [SAT] and ACT scores).

## *Discovery of Technical Errors*

Following the release of the 1994 reading data in April, 1995, an error was discovered in one of the relatively new computer programs being used to scale the results. The error involved the treatment of omitted responses in the IRT scaling of the constructed-response items that were scored using a partial-credit model. The program was written in such a way that blank responses (both omitted and not reached) were treated as missing, despite the fact that NAEP's usual procedure is to treat the two types of omitted responses differently. This error affected the results for both the 1992 and 1994 reading assessments (and the 1992 mathematics assessment), though in mostly minor ways. It did, however, necessitate a redoing and re-reporting of the results for both 1992 and 1994.

At about the same time, a second technical problem, which also affected results for both 1992 and 1994, was identified in the cutscores for the NAEP reading achievement levels. As described in chapter 6, the judges charged with setting achievement levels rated each of the assessment items in terms of the percentages of students at each level who should answer the item correctly. Information was then combined across

[15] Preliminary long-term trend results did become available just prior to the release of the 1994 *First Look* report, after the Panel's investigators had completed their research. The results showed that long-term trend data in reading dropped slightly in 1994, but the drop was not statistically significant. Furthermore, long-term trend patterns for some subgroups paralleled the findings from the short-term trend, but patterns for other subgroups were substantially different.

all judges and all items and mapped on the NAEP scale to establish the cutscores between levels. The error discovered in 1995 involved the use of an incorrectly derived formula for weighting the ratings for the items scored using a partial-credit model versus the items scored dichotomously when mapping them on the NAEP scale. As a result, the partial-credit items received much more weight than intended (three times as much as the dichotomous items), and the resultant cutscores were set too high on the reading scale.

Upon discovery of the errors, the problems were quickly corrected and the national and TSA reading data recalculated. An examination of their impact found that the errors had a minimal effect on both state and national results, and virtually no impact on policy related interpretations. A revised edition of the 1994 *First Look* reading report, based on the corrected results, was distributed, and the corrected results were used in all subsequent reading reports.[16]

The Panel recognizes that NCES and NAGB contractors make every reasonable effort at controlling the quality of their scoring, analyses, and reports. There is simply no way every procedure and every line of computer code can be double or triple checked. However, it is important that major steps be routinely and carefully checked for accuracy. In addition, the Panel recommends that NAEP contractors employed by NCES or NAGB continue to produce technical reports for each assessment and release them in a timely manner along with reports of the assessment results. The technical reports should provide thorough documentation of all design related, technical, and psychometric activities associated with the assessment, including specific item-level detail on editing and equating decisions. Although the technical reports are seldom used by the average NAEP consumer, they are a necessary prerequisite for outside review of NAEP procedures and provide an invaluable resource when questions arise about anomalies or errors.

## *Summary*

Over the years, NAEP has had an outstanding reputation for producing results that are both reliable and valid. During the 1990s, while NAEP has grown both larger and more complex, the NAEP contractor has continued to carry out the myriad of scaling and analysis activities under tight time lines and mostly without evident problems. Furthermore, NCES and ETS have implemented a number of new or revised procedures during the 1990s that appear to offer useful solutions to many of the major methodological problems encountered. At the same time, there is evidence of strain in the system. As changes continue to accumulate, the entire system must be carefully monitored on an ongoing basis to ensure its satisfactory performance.

Content area groups have frequently pushed the limits of NAEP's analytic capabilities as they introduce new frameworks and encourage the use of new item types and other assessment innovations. Policy changes also have affected various aspects for the assessment and required corresponding enhancements to the analysis design. The most current example involved modifying the 1996 administration procedures to

[16] See the revised edition of P.L. Williams, C.M. Reese, J.R. Campbell, J. Mazzeo, and G.W. Phillips, op. cit. The corrected 1992 mathematics results will be issued in conjunction with the 1996 mathematics results.

expand inclusion and permit accommodation of students with disabilities or limited English proficiency.

NAEP has also been affected by understandable pressure from NAGB to decrease the time lines for reporting. Because of the importance of the TSA data to the states and the desire to release the results for individual states as soon as possible following the administration, the contractor has made attempts to complete the scaling and analysis work more quickly. A major concern, however, is the impact on the reporting schedule when additional analyses need to be run or data reanalyzed. If problems occur, they can create an immense amount of additional work that needs to be done in a short time, all of which adds to the pressure of getting the results reported quickly and increases the likelihood that errors will creep into reported results, as they did in 1992 and 1994.

These pressures also contribute to an increasingly cumbersome and complex NAEP. The current design has been in place for more than a decade and was instituted at a time when the size, and many of the objectives, of the NAEP program were considerably different. It is quite likely that many of the current strains are the result of attempting to shore up a structure that is no longer suited to current priorities. Many others have also concluded that the time for a major design change may be upon us, and the NAEP Governing Board is currently considering broad changes that would simplify the NAEP design. ***The Panel supports NAGB's efforts to develop a new, more streamlined design for NAEP. However, the Panel also cautions that a set of principles must be established against which to evaluate alternative redesigns. Whereas reducing costs, improving the timeliness of reporting, and minimizing error are all important objectives for a new design, it is also important that the new design support all of the important objectives of the current NAEP program—including a commitment to forward-looking and innovative assessments that measure important aspects of student achievement, results that are policy relevant, and valid and reliable measurements of trend. Whichever design is decided upon, it should preserve NAEP's status as the premier indicator of students' educational progress for the nation and the states.***

In the meantime, even though the Panel's study of the fourth-grade reading assessment identified no scaling problems specific to the TSA, the Panel is concerned about the various problems that affected the 1994 NAEP analyses more generally. The Panel therefore recommends that NCES conduct or commission additional studies to validate the current analysis and scaling models. These studies should include research on the strength of the models being employed and the robustness to violations of assumptions. In addition, various procedural changes designed to improve and check the integrity of the NAEP data prior to its release should be investigated.

# 6 Reading Achievement Levels

## Introduction

In the past decade, NAEP has both influenced and been influenced by the national education reform movement, particularly by its efforts to raise students' achievement through the setting of content and performance standards. Chapter 2 of this report discussed the NAEP reading framework within the context of evaluating the content validity of the 1994 TSA. Like its counterparts in other NAEP subject areas, the reading framework is a "broad description of the knowledge and skills students should acquire in a particular subject area."[1] It therefore fits well with the general criteria for content standards. In this chapter, the Panel focuses attention on the other type of standards, referred to as performance standards, or "achievement levels" in the NAEP terminology.

In December, 1989, NAGB initiated a process to set performance standards for NAEP. If content standards define *what* American students should know and be able to do, performance standards define *how well* they should know and be able to do it. Thus, for example, a performance standard for eighth-grade writing would indicate how well an eighth-grade student should be able to write when producing the various kinds of text set out in the writing framework. According to the then NAGB Chairman Chester E. Finn, Jr., "NAEP has long had the potential not only to be descriptive but to say how good is good enough."[2]

Because of the national interest in education standards and because the achievement levels were both a new and a controversial part of the NAEP program, NCES asked the NAE Panel to undertake a special study of the achievement levels as part of its evaluation of the 1992 TSA. Based on its investigations, the Panel concluded that the achievement levels were flawed and recommended that they not be used to report the 1992 NAEP results. The Governing Board disagreed and decided to use the achievement-level reporting for both the 1992 results in mathematics and reading and the 1994 results in reading, U.S. history, and world geography.

The Panel's criticisms aside, there is little doubt of the popularity of achievement levels. Surveys of state assessment directors and curriculum specialists conducted after the release of 1992 NAEP TSA results indicated strong support for the use of achievement levels in reporting those results.[3] Furthermore, the Panel itself has indicated its belief in the value of national standards that exemplify challenging expectations for student performance, and notes that it is exactly because of their very

[1] Public Law 103-227, Section 3, 108 Stat. 129: March 1994.

[2] C.E. Finn, Jr., news release (Washington, D.C.: National Assessment Governing Board, November 29, 1989), 1.

[3] The National Academy of Education, *The Trial State Assessment: Prospects and Realities* (Stanford, CA: Author, 1993), xxiv.

popularity and potential utility that the quality of the achievement levels is so important. Drawing on its *quality principle*, the Panel therefore repeats a key statement made in the executive summary for its second evaluation report, *Setting Performance Standards for Student Achievement:* "The standards set must be defensible in order to ensure that assessment data and national education policy based on the standards are sound."[4]

This chapter revisits the achievement levels in light of the Panel's earlier critique and of NAGB's response to the issues it raised. Two things should be noted from the outset. First, this chapter differs from others in this volume because it contains relatively few new analyses; instead it draws substantially from work done in the aforementioned report, *Setting Performance Standards for Student Achievement*. Second, the analyses reported in this chapter are not part of a formal evaluation of the U.S. history and world geography achievement levels set in 1994—the Panel was not asked to undertake such an evaluation. Nevertheless, because NAGB had commissioned work to respond to the Panel's earlier criticisms and because the 1994 reading results were reported using the achievement levels, the Panel determined that this evaluation of the TSA would be incomplete without consideration of this important component. *It is the Panel's view that the achievement levels, as they currently exist, continue to have fundamental problems and that it is therefore responsible and appropriate to question their continued use for reporting NAEP results.*

The chapter begins with a brief overview of the process by which the achievement levels were set in 1992. The work that the Panel undertook to evaluate the achievement levels is summarized next, and this is followed by a review of the "revisitation" study commissioned by NAGB to respond to the Panel's criticisms. The achievement-level results from the 1994 U.S. history and world geography assessments are also commented on briefly because, although not directly applicable to the TSA, they illustrate that several issues raised about the achievement levels in 1992 remain unresolved. The chapter ends with a Panel recommendation on the continued use of the current achievement levels.

## *The Setting of Performance Standards on NAEP*

After determining in 1989 that it would set achievement levels for NAEP, the Governing Board's first step was to select a method for doing so. The testing and measurement literature documents a large number of standard-setting methods which generally can be grouped into two large categories: those that are test-centered in their approach and those that are examinee-centered.[5] After considering various alternatives, NAGB selected the Angoff procedure, a commonly used test-centered model for setting a cutscore on an existing test.

[4] The National Academy of Education, *Setting Performance Standards for Student Achievement* (Stanford, CA: Author, 1993), xxiv.

[5] R.M. Jaeger, "Certification of student competence," in *Educational Measurement: Third Edition* (New York: American Council on Education and Macmillan Publishing Company, 1993), 493-497.

The Governing Board's adoption of generic definitions of basic, proficient, and advanced performance levels was the next step in the level-setting process. These generic definitions, presented in figure 6.1, served as the basis for the development of more specific and substantive descriptions of the three levels in each NAEP subject area and at each grade assessed—4, 8, and 12.[6] In keeping with the modified Angoff procedure,[7] the expert judges selected to set the achievement levels used these subject specific descriptions to formulate mental pictures of students who would minimally meet the implied performance criteria for each level. The judges then reviewed individual NAEP items and estimated the percentages of these hypothetical students who would be likely to answer each item correctly.[8] These item-by-item judgments were converted to the NAEP scale based on each item's difficulty and related characteristics. Information was then averaged across items and all panelists to arrive at the cutscores that would distinguish four levels of achievement at each grade level. That is, at a given grade level, students whose NAEP scores (actually their plausible values, see chapter 5) fell below the lowest cutscore would have their performance classified as "below basic." Those with higher scores would be considered to have performed at the basic, proficient, or advanced level, depending upon which of the cutscores they surpassed.

**Figure 6.1.** ***Achievement-level definitions adopted by NAGB on May 11, 1990***

**Basic**

Denotes partial mastery of the knowledge and skills that are fundamental for proficient work at each grade—4, 8, and 12. For 12th grade this is higher than minimum competency skills (which normally are taught in elementary and junior high schools) and covers significant elements of standard high school-level work.

**Proficient**

Represents solid academic performance for each grade tested—4, 8, and 12—and reflects a consensus that students reaching such a level have demonstrated competency over challenging subject matter and are well prepared for the next level of schooling. At grade 12, the proficient level will encompass a body of subject-matter knowledge and analytical skills, of cultural literacy and insight, that all high school graduates should have for democratic citizenship, responsible adulthood, and productive work.

**Advanced**

Signifies superior performance beyond proficient grade-level mastery at grades 4, 8, and 12. For 12th grade the advanced level shows readiness for rigorous college courses, advanced technical training, or employment requiring advanced academic achievement. As data become available, it may be based in part on international comparisons of academic achievement and may also be related to Advanced Placement (AP) and other college placement exams.

[6] The generic definitions in figure 6.1 were in place when the Panel did its evaluation of the 1992 achievement levels. Since then, the generic descriptions have been revised.

[7] NAGB modified the procedure to allow for several rather than a single cutscore, as is usually the case where the Angoff method is used.

[8] For partial-credit items, in which students received scores based on the quality of their responses rather than simply being scored as "right" or "wrong," somewhat different rating methods were used.

## Criticisms of the NAEP Achievement Levels

Because NAEP achievement levels were the first highly visible example of the development of national performance standards, NAGB's standard-setting procedure was the focus of considerable attention and several independent evaluations. Evaluations conducted by Linn et al. for the NAEP Technical Review Panel and by Stufflebeam et al. were critical of the achievement levels used to report the 1990 mathematics results.[9] The U.S. Government Accounting Office (GAO) also issued critical evaluations of the 1990 levels and of the new reading and mathematics levels developed for 1992.[10] Overall, the criticisms included the following: the judgment tasks required by the modified Angoff process were found to be difficult and confusing; the NAEP item pool was not adequate to reliably estimate performance at the advanced levels; appropriate validity evidence for the cutscores was lacking; and neither the descriptions of student competencies nor the exemplar items selected were appropriate for describing actual student performance at the achievement levels defined by the cutscores. These findings engendered substantial discussion and debate regarding the use of the achievement levels to support inferences about levels of student ability.

In light of the controversy surrounding the evaluations of the 1990 achievement levels and the increasing interest in education standards, NCES asked the NAE Panel to conduct an additional independent evaluation of the 1992 achievement levels in reading and mathematics. The Panel's focus was on the adequacy of the achievement levels and whether the levels could lead to valid inferences about levels of student performance.[11] The evaluation addressed two major questions:

- Are the processes used to set the levels internally consistent and coherent?
- Do the final levels appear to be reasonable and valid on the basis of comparisons to external data and substantive analyses?

The results of the Panel's evaluation raised serious questions about

- The adequacy of the Angoff or any other item-by-item method for setting achievement levels;

[9] R.L. Linn, D.M. Koretz, E.L. Baker, and L. Burstein, *The Validity and Credibility of the Achievement Levels for the 1990 National Assessment of Educational Progress in Mathematics* (Los Angeles, CA: Center for Research on Evaluation, Standards, and Student Testing, January 1991); D.M. Stufflebeam, R.M. Jaeger, and M. Scriven, *Summative Evaluation of the National Assessment Governing Board's Inaugural 1990-91 Effort to Set Achievement Levels on the National Assessment of Educational Progress* (Washington, D.C.: National Assessment Governing Board, 1991).

[10] U.S. General Accounting Office, *National Assessment Technical Quality,* GAO/PEMD-92-22R (Washington, D.C.: Author, March 1992); U.S. General Accounting Office, *Educational Achievement Standards: NAGB's Approach Yields Misleading Interpretations,* GAO/PEMD-93-12 (Washington, D.C.: Author, June 1993).

[11] For a complete description of the evaluation and its findings see *Setting Performance Standards for Student Achievement* (Stanford, CA: The National Academy of Education, 1993).

- The reasonableness of the student performance distributions established by the achievement-level cutscores, especially as compared to the performance distributions obtained from various non-NAEP measures; and
- The match between the verbal descriptions of the achievement levels and the types of items that children who scored at those levels actually could do.

The Panel further concluded that interpreting NAEP results based on invalid and unreliable achievement levels could actually harm the credibility of NAEP and jeopardize other national efforts to develop content and performance standards. A summary of some of the key findings from the various evaluation studies conducted by the Panel follows.

## *Findings from the Panel's Evaluation of the 1992 Mathematics and Reading Achievement Levels*

### *Internal Consistency*

The Panel found large internal inconsistencies in judges' ratings due to the impact of item features that should have been irrelevant for cutscore determination. That is, judges' ratings appeared to be affected by whether items were easy or hard, scored dichotomously (right/wrong) or scored on a graded scale that allowed partial credit, and, among dichotomously-scored items, whether the items were multiple choice or constructed response.[12] For example, if the fourth-grade basic proficiency level was determined using the judges' ratings on dichotomous items only, the result would be a cutscore of 191. (See table 6.1). By comparison, if only partial-credit items were used, the resulting cutscore for fourth-grade basic proficiency would be 281. This is a difference of 90 points in cutscores for the *same proficiency level* at the *same grade*. By contrast, the differences are much smaller when one considers cutscores set for different grades but using the same item type. Thus, if only dichotomous items were used, the cross-grade differences in cutscores for the basic level would have been only 43 points between 4th and 8th grades, and only 56 points between 4th and 12th grades. Note that the 90-point difference between the two cutscores for 4th-grade basic exceeds the 4th- to 12th-grade difference by 34 points, or more than 60 percent. The data in table 6.1 reveal a similar pattern for the other proficiency levels and grades. Overall, cutscores set using dichotomous versus partial-credit items differed by as much as, or more than, the cutscores set for adjacent grades at the same achievement level using dichotomous items only. This was true *even though the panelists' judgments had been adjusted for differences in item difficulty before*

[12] In 1992, dichotomously-scored items included both multiple-choice items and short-answer constructed-response items. Partial-credit items were also constructed response but required a longer, or extended, student response.

*computing the cutscores*. The Panel found similar internal inconsistencies for multiple-choice versus dichotomously-scored constructed-response questions and easy versus difficult questions.

**Table 6.1.** ***Reading achievement-level cutscores, separately for dichotomous items and partial-credit items***

| Achievement Levels/Item Types | Grade 4 | Grade 8 | Grade 12 |
|---|---|---|---|
| *Basic* | | | |
| Dichotomous Items | 191 | 233 | 250 |
| Partial-Credit Items | 281 | 290 | 329 |
| *Proficient* | | | |
| Dichotomous Items | 230 | 272 | 294 |
| Partial-Credit Items | 317 | 336 | 363 |
| *Advanced* | | | |
| Dichotomous Items | 260 | 311 | 337 |
| Partial-Credit Items | 356 | 389 | 394 |

SOURCE: D.H. McLaughlin, "Validity of the 1992 NAEP Achievement Level-Setting Process," in *Setting Performance Standards for Student Achievement: Background Studies* (Stanford, CA: The National Academy of Education, 1993), 88.

The internal consistency problems are important and especially troubling because of the implications for comparing NAEP results over time. By way of example, consider again the impact of dichotomous versus partial-credit items. In recent years, the pattern has been to change the mix of these items over time so that more partial-credit items are included on each new assessment (e.g., the 1990, 1992, and 1996 mathematics assessments). Because the different item types generate such different achievement-level cutscores however, such changes in the mix will change the implications of the cutscores and generate achievement-level results that do not reflect the same conception of advanced, proficient, and basic performance in successive years.

Based on the findings of internal inconsistencies, the Panel concluded *that standard-setting judges were unable to maintain a consistent view of their expectations regarding basic, proficient, and advanced performance. In fact, the Panel concluded that the cognitive judgments that were required of the judges—holding a consistent mental image of students who minimally meet the requirements of each achievement level and then determining the percentage of such children who would respond correctly to each of the items in the assessment—were virtually impossible to make. As a result, the Panel judged that the procedure used in setting the 1992 reading and mathematics achievement levels was seriously flawed.*

The work done by Thorndike further supports the Panel's conclusion. Thorndike found that, although judges are able to rank order items in terms of their relative difficulty (i.e., in terms of the relative proportion of students who would answer each item correctly), the judges cannot accurately guess the specific percentage of students

who will answer an item correctly.[13] Unfortunately, it is this latter skill that is required of the judges when they set the achievement levels by means of the Angoff procedure. Dissecting the problem even more, McLaughlin, in a letter to an NCES official, pointed out "to estimate the percentage of students at a level of performance who would answer an item correctly, [judges] would need to delineate the ways that students could answer an item, to relate these to cognitive processes that students may or may not possess, and operationally link these processes with the categorization of performance."[14] McLaughlin went on to say that the near impossibility of these inter-related tasks manifests itself in the fact that judges were opting for an easier task (i.e., of merely generating reasonable-sounding numbers that reflected the differences in item difficulty that they were given as background information).

## *The External Comparison Studies*

The Panel also conducted a number of external comparison studies to assess the reasonableness of the cutscores that were produced. Among these comparisons were classroom based reading and mathematics studies at grades four and eight. Each of the four studies (counting each subject and grade as a separate study) involved more than 100 students who also completed NAEP assessments as part of the 1993 field test. In three of the four studies, teachers' ratings and individual assessments administered by researchers consistently found more students performing at the advanced, proficient, and basic levels than were identified in these categories by applying the achievement-level cutscores to their performances on NAEP.[15] For example, the data in table 6.2 show that teacher judgments identified 11 percent of the fourth-grade students as performing at or above the advanced level in reading, and the researchers identified 15 percent. In contrast, the NAEP achievement-level cutscores placed only 0.7 percent of these same students at or above the advanced level. Similar discrepancies were found between teacher ratings and classifications based on the NAEP cutscores in eighth-grade reading and in fourth- and eighth-grade mathematics.

In addition, at grade 12, data from the SAT and AP examinations suggested that there were more advanced students in reading and mathematics than were found to be advanced by the 12th-grade achievement levels. For example, comparisons showed that in 1992, 5.8 percent of students scored at a high level (550 or above out of 800) on the SAT Verbal test, which measures vocabulary, verbal reasoning, and reading comprehension, whereas only 3.2 percent of U.S. 12th graders scored at the advanced level in reading using the NAEP achievement levels. Similarly, with regard to mathematics, only 2 percent of U.S. 12th graders scored at the advanced level on

[13] R.L. Thorndike, "Item and Score Conversion by Pooled Judgment," in *Test Equating*, ed. P.W. Holland and D.B. Rubin (New York: Academic Press, 1982), 309-317.

[14] D.H. McLaughlin, *The Problem with Item-Based Judgment Procedures for NAEP Achievement Level Setting* (January 1994), a draft document presented to Dr. John Burkett, NCES, January 25, 1994.

[15] To rate their students, teachers used the NAEP achievement-levels descriptions of what students performing at given achievement levels should know and be able to do.

**Table 6.2.** ***Percentages of fourth-grade field-test students achieving basic, proficient, and advanced reading performance based on three sources***

| Achievement Level | Criterion based on NAEP performance and cutscores set by **Governing Board** | | Criterion based on **teachers'** ratings of classroom performance | | Criterion based on **researchers'** ratings of individual assessment performance | |
|---|---|---|---|---|---|---|
| | Cutscore | Percent Scoring at or above | Cutscore | Percent Scoring at or above | Cutscore | Percent Scoring at or above |
| Basic | 212 | 42 | 171 | 80 | 188 | 68 |
| Proficient | 243 | 11 | 214 | 39 | 226 | 30 |
| Advanced | 275 | 0.7 | 249 | 11 | 244 | 15 |

SOURCE: The National Academy of Education, *Setting Performance Standards for Student Achievement* (Stanford, CA: Author, 1993), 85.

NAEP while at least 7.5 percent of high school graduates scored 600 or better on the SAT Mathematics test.[16, 17]

Based on the overall findings of its evaluation of the 1992 achievement levels, the Panel recommended that NAGB discontinue the use of the Angoff method and urged NCES and NAGB not to report the 1992 NAEP results using the achievement levels. It is important to restate, however, that despite its criticism of the achievement levels, the Panel, both then and now, endorses the use of performance standards as a way of reporting NAEP results. The Panel's statement in its report on the evaluation of the 1992 achievement levels still stands:

> The members of the NAE Panel strongly affirm the potential value of performance standards and test-score achievement levels when linked to comprehensive content standards as a way to establish clear expectations for teachers, students, and the public. Furthermore, we strongly affirm the value of performance standards

[16] The 1992 SAT Verbal test was known to be much more difficult than the SAT Mathematics test. The College Board subsequently rescaled the Verbal test to make the score scales for the two measures more equivalent. Because of the known differences in difficulty, the Panel adopted the cutpoint of 550 on the Verbal test as roughly equivalent to 600 on the Mathematics test. For a more detailed description of the methodology and findings for all of the external comparison studies see *Setting Performance Standards for Student Achievement* (Stanford, CA: The National Academy of Education, 1993).

[17] In response to criticism that no research had been done to verify that advanced performance at the 12th grade was related to readiness for rigorous college courses, advanced technical training, or employment requiring advanced academic achievement, the Governing Board removed this language from the revised description of advanced achievement.

as one method for NAEP reporting, both to improve the interpretableness and usefulness of the Nation's Report Card and to improve information to further the nation's education improvement.[18]

## *NAGB's Revisitation of the 1992 Achievement Levels*

As noted, NAGB decided to use the achievement levels for reporting the 1992 results, the Panel's findings and recommendations notwithstanding. The Governing Board did agree, however, to conduct an additional study of the 1992 reading achievement levels before deciding whether to use the levels to report the results of the 1994 reading assessment. This "revisitation study," as they called it, could have been an opportunity to address some of the substantive issues that the Panel raised in its report. Unfortunately, the design of the NAGB study focused solely on the Panel's finding of a mismatch between what students know and can do at the various levels and how the achievement levels are described. It did not address the other important issues the Panel raised in its evaluation of the achievement-levels setting process. Most critical was the study's failure to address the internal inconsistencies in judges' ratings by item type, as discussed above.

### *Overview of the Revisitation Study*

The revisitation[19] study used two procedures to examine whether there was a disjuncture between the achievement levels actually used by NAGB and the descriptions implied by items at or below the cutscores. In the first procedure, called the judgmental item-categorization procedure, a panel of experts was asked to determine whether each of the items could be placed into an achievement-level category based on the achievement-levels descriptions. They were then asked to assess whether the NAEP items adequately represented the knowledge and skills described in the achievement-level descriptions. In the second procedure, called the item difficulty-categorization procedure, a second panel of grade-level experts was asked to examine how well students performing at a given achievement level did on a range of items and to consider whether their performance on these items accurately reflected the knowledge and skills that the achievement-level description indicates (i.e., can students performing at specific achievement levels do what the description say they should be able to do).

[18] The National Academy of Education, *Setting Performance Standards for Student Achievement* (Stanford, CA: Author, 1993), 148.

[19] This review of the revisitation study is based on information provided in the American College Testing Program, *NAEP Reading Revisit: An Evaluation of the 1992 Achievement Levels Descriptions* (Author, February 1995).

The two panels of experts were subsequently brought together and asked for a summative judgment: after participating in the study, did they believe the achievement-level descriptions should or should not be used for reporting the 1994 NAEP reading assessment results? The panelists concluded that, at all three grade levels, students who scored at the given achievement levels knew and were generally able to do what the achievement-level descriptions indicated they should know and be able to do, and they therefore supported the continuing use of the achievement-level descriptions. They did recommend some format and editorial changes however, in addition to small substantive revisions. For example, the panelists thought that fourth graders at the basic level were able to draw inferences from text and that this should be added to the basic definition at the fourth grade.[20]

### *Sorting Items into Achievement Categories Does Not Confirm Specific Cutscores*

The NAE Evaluation Panel has serious reservations about the underlying research question and design of the NAGB revisitation study, and thus about the conclusion reached by its participating experts. First of all, it is important to keep in mind that the achievement-levels descriptions are quite general in nature and the factors distinguishing one level from another are not particularly distinct. For example, at the fourth grade, students performing at the basic level should, among other things, be able to "make relatively obvious connections between the text and their own experience," and "extend the ideas in the text by making simple inferences." The same behaviors are required for performance at the proficient level except that the qualifiers "relatively obvious" and "simple" are dropped. Furthermore, some of the behaviors distinguishing the levels (e.g., "Demonstrate an awareness of how authors compose and use literary devices" at the fourth-grade advanced level) would not be applicable to the majority of items.[21] Thus, the finding of an adequate match between the descriptions and judges' ability to rank items into the relevant levels is not particularly surprising because, for most items, the panelists need simply rank the items by difficulty. Drawing from the work of Thorndike, which indicates that judges

[20] The detailed recommendations for changes in definitions for all three grade levels are given in the American College Testing Program, op. cit., 18-20.

[21] The complete reading achievement-level descriptions for grade four state that "Fourth-grade students performing at the Basic level should demonstrate an understanding of the overall meaning of what they read. When reading text appropriate for fourth graders, they should be able to make relatively obvious connections between the text and their own experiences, and extend the ideas in the text by making simple inferences. Fourth-grade students performing at the Proficient level should be able to demonstrate an overall understanding of the text, providing inferential as well as literal information. When reading text appropriate to fourth grade, they should be able to extend the ideas in the text by making inferences, drawing conclusions, and making connections to their own experiences. The connection between the text and what the student infers should be clear. Fourth-grade students performing at the Advanced level should be able to generalize about topics in the reading selection and demonstrate an awareness of how authors compose and use literary devices. When reading text appropriate to fourth grade, they should be able to judge texts critically and, in general, give thorough answers that indicate careful thought." J.R. Campbell, P.L. Donahue, C.M. Reese, and G.W. Phillips, *NAEP 1994 Reading Report Card for the Nation and the States* (Washington, D.C.: National Center for Education Statistics, January 1996), 42.

are able to estimate fairly accurately the relative difficulty ranking of items,[22] we would expect to find that the panelists were able to map with reasonable accuracy one set of rankings (the item difficulties) onto a second set of rankings, the achievement levels. However, the relationship between such ranking and the determination of a numerical cutscore is hardly straightforward.

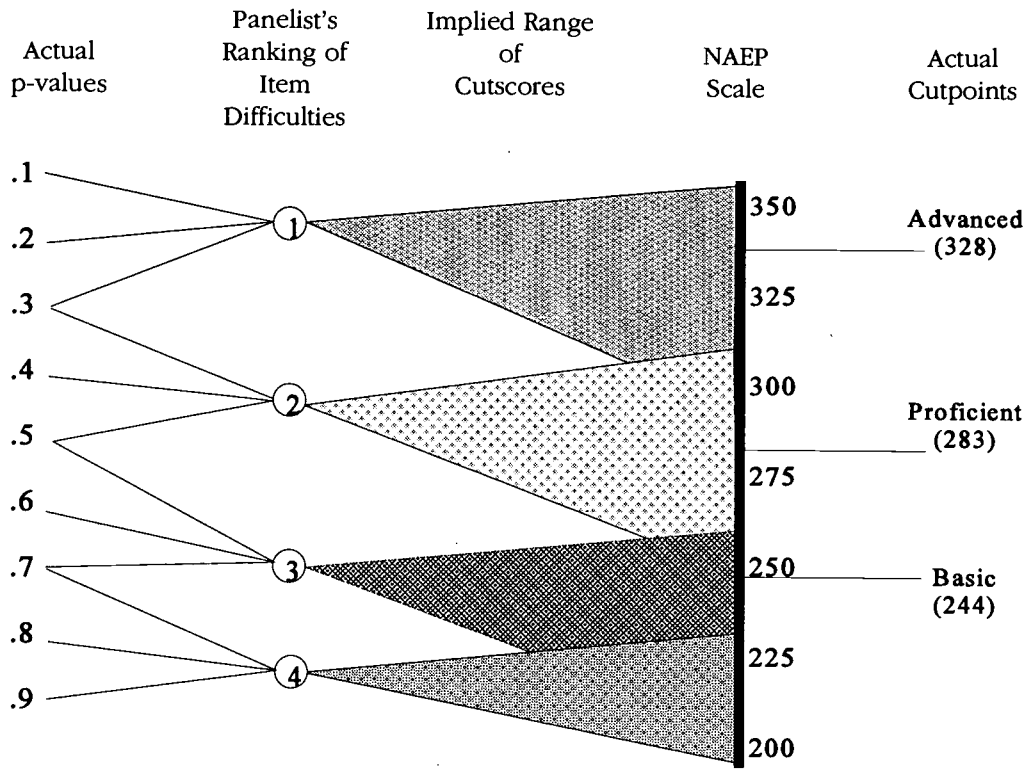
Figure 6.2 illustrates this point. It portrays a hypothetical, idealized example of the relationship between item p-values (where p-value is the percentage of examinees who actually answered the item correctly), a panelist's item rankings, and the implied range of achievement-level cutscores. The hypothetical panelist depicted here has conceived of four ranked clusters of items that, in this case, align reasonably well with the achievement-level categories. Although she does not know the actual p-values of the items, she is able to make reasonable inferences about their relative difficulty based on her reading of the item text. Thus she has sorted all of the items that were answered correctly by only 10 or 20 percent of the students (those with actual p-values of .1 or .2) into the highest ranked, or "hardest," cluster. Similarly, all of the items with p-values of .4 have been sorted into the second cluster. However, among the items that had p-values of .3, she has sorted some into the highest cluster and some into the second. The same pattern is seen for the rest of her sorts and is consistent with Thorndike's assertion that panelists are able to rank items with reasonable, but not perfect, reliability.

Recall that the achievement-level cutscores were originally set on the basis of the actual difficulties of the items that judges indicated should be answered correctly by students who minimally met the standards for each achievement level. Therefore, because our hypothetical panelist has done a reasonably good job of estimating actual item difficulty, and because she also holds a standard for student performance that is at least somewhat like that of the original judges, she is able to roughly align her ranked clusters with the official achievement-level categories. However, her sort, even under these idealized conditions, is too crude to align with any *specific* cutscore on the NAEP scale. At best, the ranked clusters can only be aligned with a *range* of cutscores (indicated by the overlapping projections onto the scale at the right-hand side of figure 6.2). Because the panelists' judgments do not translate into specific cutscores, the revisitation study is unable to address the issue of whether or not the cutscores on the 1992 NAEP reading assessment were too high.

Again, figure 6.2 is a highly idealized model for one hypothetical panelist. Other panelists might begin by sorting the items in seven, eight, or nine clusters before translating their rankings into decisions about which achievement level to associate with each item. Moreover, different panelists will have been more or less exact in their original item sorts; the less that their sorts match the true difficulty rankings of the items, the wider the range of possible cutscores that could be implied by their decision.

[22] R.L. Thorndike, op. cit., 310.

**Figure 6.2.** ***An illustration of the relationship between actual item p-values, one panelist's ranking of item difficulties, and the implied range of cutscores for achievement levels.***




Because the achievement-level descriptions are so general, and because items ranked by relative difficulty will tend to align with the achievement levels, the Panel expressed concern about what it saw as a "confirmatory bias" in the revisitation study from the time when it was originally designed. More importantly, as argued above, the study could not in any case address the Panel's perception that the cutscores may have been set too high. The Panel would have preferred a study that better allowed for the disconfirmation of the hypothesis that there was a good fit between actual performance and the performance implied by the achievement levels. As Cronbach has pointed out, the goal of conducting validity studies is to maximize the opportunity for disconfirmation rather than maximizing the chance of confirmation.[23]

## *Statistical Evidence from the Revisitation Study*

Following the session with the two sets of panelists, the NAGB contractor undertook a series of statistical analyses, based on the data provided by the panelists, to further

[23] L.J. Cronbach, "Construct validation after thirty years," *Intelligence: Measurement, Theory, and Practice*, ed. R.L. Linn (Urbana, IL: University of Illinois Press, 1989), 147-171.

examine the fit between the achievement-levels descriptions and what students know and actually do. Although the results of these analyses were intended to provide additional support for the conclusions of the expert panelists, the Panel, upon reviewing the data, found the evidence less than compelling. For example, recall that experts on one of the panels were asked to read the knowledge and skills described in the achievement-levels descriptions and then to categorize items according to the level of knowledge and skills—basic, proficient, or advanced—that they believed were required to get the item correct. For the statistical analysis, the contractor reviewed data on the actual percentage of students who answered each item correctly, then computed the average and the highest and lowest percent correct for the items that the panelists had classified at each achievement level. These data are presented by achievement-level category and grade in table 6.3.

**Table 6.3.** ***Average percent correct of items categorized by judges at each achievement level by grade*** (numbers in the brackets are the lowest and highest percent correct for items in that cell)

| Judges' Categorization of items | Grade 4 | Grade 8 | Grade 12 |
|---|---|---|---|
| Basic | 58<br>[15 - 93] | 59<br>[4 - 91] | 69<br>[12 - 97] |
| Proficient | 54<br>[3-90] | 31<br>[2 - 77] | 56<br>[6 - 94] |
| Advanced | 32<br>[0 - 78] | 2<br>[0 - 5] | 21<br>[1 - 82] |

SOURCE: American College Testing Program, *NAEP Reading Revisit: An Evaluation of the 1992 Achievement Levels Descriptions* (Author, February 1995), 14.

The average percents correct were then examined across achievement levels to see whether they followed the expected patterns. For example, one would expect that fourth graders would correctly answer a higher percentage of items categorized as basic than they would items categorized as proficient or advanced. In fact, at each grade level, the pattern of percent correct did match expectations. For example, at grade four, the percent correct for items categorized at the basic level averaged 58 percent; for items at the proficient level, 54; and for items categorized at the advanced level, 32. The average percent correct was therefore higher for items categorized at the basic level than for those categorized at the proficient level, and was higher for items categorized at the proficient level than it was for those categorized at the advanced level. This pattern held true for all grade levels.

As can also be seen in table 6.3 however, whereas the *averaged* percents correct behaved as expected, there was tremendous variation of percent correct within each category. Thus, the fourth-grade items categorized at the basic level had percents

correct that ranged from 15 percent to 93 percent. This wide range of percents correct in each group of items across all grades and achievement levels weakens the evidence provided by simply looking at the averages.

## *Analysis of the Reported Hit Rate*

The contractor next conducted rather complex analyses of "hit rates."[24] To understand the concept of a hit rate, note that both students and items can be categorized into achievement levels. In the revisitation study, students were so classified on the basis of their NAEP score,[25] and items were so classified based on the judgments of the expert judges. The assumption is that students who perform at a given achievement level should generally be able to respond correctly to items categorized at or below that level and should generally not be able to answer items classified at a higher level. When actual NAEP results confirmed this pattern for a given item, the analyst labeled it a "hit." When the results disconfirmed the expected pattern, the item was labeled a "miss." The hit rate is the number of hits divided by the total number of items at each respective grade level.

The contractor performed nine analyses of hit rates (one for each of three achievement levels at three grade levels) and found hit rates from 48 percent to 97 percent with a median hit rate of 81 percent across the nine analyses. Although these rates seem impressive, further consideration diminishes their persuasiveness. The Panel found that the marginal distributions of the tables were such that the hit rates on the basis of chance alone ranged from 44 percent to 93 percent with a median of 72 percent. Thus, the number of hits was not far above what one would expect to observe on the basis of chance.

In conclusion, the statistical data from the revisitation study provide no more than modest evidence in support of a positive relationship between the achievement-level descriptions and what students know and can do. More importantly however, the analyses undertaken were incapable of addressing a question of central importance to the Panel—were the levels set too high or not?

## *Summary*

*The Panel's review of both the qualitative and the quantitative data compiled by NAGB concluded that neither set of results from the revisitation study dealt with one of the Panel's central concerns: whether the cutscores for the achievement levels were or were*

[24] Details of the hit rate analyses can be found in the American College Testing Program, op. cit. The Panel's full analysis of the revisitation study is presented in G.W. Bohrnstedt and E. Hawkins, "Reporting the 1994 Reading Results by Achievement Levels," *Quality and Utility: The 1994 Trial State Assessment in Reading, Background Studies* (Stanford, CA: The National Academy of Education, forthcoming).

[25] Note that this discussion is simplified for illustrative purposes; as described in chapter 5, NAEP does not actually produce individual student scores.

*not too high. Furthermore, NAGB commissioned no studies designed to address the Panel's conclusion that the procedure for establishing achievement levels had generated serious* ***internal*** *inconsistencies.*

## *Continuing Issues regarding the Achievement-Levels Setting Procedure*

Because U.S. history and world geography were administered only at the national level in 1994, the Panel was not charged with evaluating the achievement-levels setting process for these subjects. Nonetheless, the Panel examined information published about them in order to obtain a more complete context for evaluating the reliability of the item-by-item achievement-levels setting process that NAGB continues to employ. In its review of data from the standard-setting meetings for U.S. history and world geography, the Panel still found that the internal inconsistencies (as a function of item type) remained, albeit at a somewhat diminished level. Despite procedural modifications made by the contractor, differences of 25 to 30 points between levels set by the dichotomous versus the partial-credit items for these subjects were not unusual and in every case, the levels set by the partial-credit items were higher. (See table 6.4.) Furthermore, although not mentioned in the release of either the U.S. history or world geography results, a comparison of results for these two subjects raises additional questions of external validity for the achievement levels. More specifically, only 12 percent of 12th-grade students were classified as at or above the proficient level in U.S. history, but 29 percent were so classified in geography. These results are very counterintuitive because virtually every high school student in the nation takes U.S. history and very few take geography. Indeed, the results were so puzzling that the Governing Board seriously considered not releasing the 12th-grade results for history.

In *Setting Performance Standards for the Nation,* the Panel used the percentage of students receiving a score of three or more on the AP examination as a useful, although admittedly imperfect way to determine whether the cutpoints in the 1992 reading and mathematics assessments had been set too high for the advanced level.[26] The AP program, operated under the auspices of the College Board, provides curricular materials to participating high schools that allow the schools to offer college-level courses to high school students. High school seniors who take the AP courses then have the option of taking AP examinations which are scored from one to five, five being the highest possible score. Many colleges and universities award college credit for the courses if the student's scores are three or higher. The evidence from the distributions of student AP scores corroborated the Panel's hypothesis that the cutscores at the 12th-grade advanced level were set too high in the 1992 NAEP mathematics and reading assessments.

Similar analyses were undertaken for U.S. history. (No AP program is available for world geography.) The distribution of U.S. history AP scores for 1994 graduating

[26] The National Academy of Education, *Setting Performance Standards for Student Achievement* (Stanford, CA: Author, 1993), 92-93.

**Table 6.4.** ***Achievement-level cutscores for U.S. history and world geography, separately for dichotomous and partial-credit items***

| Achievement Levels/Item Types | Grade 4 | Grade 8 | Grade 12 |
|---|---|---|---|
| World Geography | | | |
| *Basic* | | | |
| Dichotomous Items | 182 | 230 | 243 |
| Partial-Credit Items | 188 | 247 | 272 |
| *Proficient* | | | |
| Dichotomous Items | 236 | 275 | 295 |
| Partial-Credit Items | 244 | 291 | 313 |
| *Advanced* | | | |
| Dichotomous Items | 271 | 306 | 329 |
| Partial-Credit Items | 286 | 330 | 350 |
| U.S. History | | | |
| *Basic* | | | |
| Dichotomous Items | 171 | 226 | 264 |
| Partial-Credit Items | 200 | 261 | 303 |
| *Proficient* | | | |
| Dichotomous Items | 239 | 282 | 315 |
| Partial-Credit Items | 246 | 302 | 334 |
| *Advanced* | | | |
| Dichotomous Items | 272 | 321 | 346 |
| Partial-Credit Items | 283 | 334 | 365 |

SOURCE: Communication from the National Center for Education Statistics.

seniors is shown in table 6.5. The number of students at each score level can be found in the second column. These numbers are translated into percentages of all high school graduates (the number of AP examinees at each level divided by an estimate of the total number of 1994 high school graduates) in the third column, and the cumulative percent scoring three or better is given in the last column.

The NAEP achievement levels for U.S. history estimated that, in 1994, 1 percent of the nation's seniors were at the advanced level or better. By comparison, the AP results show that 2.8 percent qualified for college credit in U.S. history, suggesting at least advanced performance on their part.[27] It should be pointed out that *the AP percentage is almost certainly an underestimate* of the number of students who would

[27] In fact, scoring three or better on the AP examination might be considered "better than advanced" because it signifies *completion* of college-level work rather than being prepared for college work, the latter being part of the original generic definition of advanced performance on the achievement levels.

pass an AP course in U.S. history if AP programs were available in all U.S. secondary schools instead of less than half of them. These results suggest that the cutscores at the advance level on the 1994 NAEP U.S. history assessment are too high, and support the Panel's conclusion that the 12th-grade cutscores at the advanced level on the 1994 NAEP reading and mathematics assessments were also very likely set too high. *These findings are important because they provide additional evidence that the Governing Board's achievement levels are identifying too few 12th-grade students as performing at the advanced level.*

**Table 6.5.** ***The 1994 Advanced Placement test results in U.S. history by score level and as a percentage of all high school graduates***

| AP Score | U.S. History | Total as a Percentage of High School Graduates | Cumulative Percentage Scoring 3 or Higher |
|---|---|---|---|
| 1 | 11,845 | 0.5 | |
| 2 | 44,275 | 1.8 | |
| 3 | 31,974 | 1.3 | 1.3 |
| 4 | 24,169 | 1.0 | 2.4 |
| 5 | 10,832 | 0.4 | 2.8 |
| Total | 123,095 | 5.0 | |

SOURCES: College Board and U.S. Census Bureau

## *Summary and Recommendations for the Use of Achievement Levels*

When one looks at the history of the achievement levels since they were first used on a trial basis in 1990, their developmental nature is highlighted. The 1990 mathematics results were found by both the Technical Review Panel and by NAGB's own evaluation team to be badly flawed. The GAO looked at the 1990 mathematics and 1992 mathematics and reading achievement levels and, like the NAE Panel, found them seriously wanting. Both the GAO and the NAE Panel recommended that the achievement levels not be used for reporting NAEP results.[28] As reported above, the achievement levels for the 1994 U.S. history and world geography assessments also raise serious questions. For example, at least some of the problems with internal consistency that the NAE Panel found in 1992 appeared again in these 1994 assessments. Additionally, the comparative percentages of students at the proficient level or above for U.S. history and world geography are counterintuitive given course

[28] NAGB had intended to report the 1992 NAEP writing results using achievement levels. However, the writing assessment scaled so poorly that NAGB, to its credit, decided against doing so.

taking patterns for the two subject areas. Finally, the results from the 1994 U.S. history AP examination again suggests that the achievement levels are identifying too few 12th-grade students as performing at the advanced level.

Combined, this evidence led the Panel to conclude that the results reported by the current achievement levels lack adequate validity and are potentially misleading. As the Panel stated in its report, *Setting Performance Standards for Student Achievement,* the standards set must be defensible in order to ensure that assessment data and national education policy based on the standards are sound.

The Panel commends NAGB for its pioneering efforts to implement performance standards for student achievement—it has been a bold experiment. However, because of fundamental problems with internal inconsistencies for the 1990 and 1992 mathematics and the 1992 reading achievement levels, continuing internal consistency problems with the 1994 history and world geography achievement levels, and apparent invalidities in the 1992 reading and mathematics achievement levels when measured against external evidence of student achievement, the Panel continues to believe that item-by-item methods for setting performance standards are fundamentally flawed as applied to student assessments such as NAEP. The central problem is the nearly impossible cognitive task of estimating the probability that a hypothetical student at the boundary of a given achievement level will get a particular item correct.

As mentioned at the beginning of this chapter, the Panel is aware that reporting by achievement levels is popular with NAEP consumers and that the states have endorsed their use.[29] Although the Panel continues to have serious reservations about the quality of the current achievement levels, it also respects the requests for reporting by performance standards and recognizes that the discontinuation of reporting by means of the current achievement levels would undoubtedly cause confusion, frustration, and dismay among many of the current consumers of NAEP results. For this reason, the Panel understands why NAGB might decide to continue to use the current achievement levels, the Panel's concerns notwithstanding. The Panel makes the following recommendations. First, NAGB should institute a competition for the design of new methods for setting performance standards for all NAEP subjects with the goal of having a new methodology in place by the time of the year 2000 NAEP assessment. Second, as recommended in the Panel's previous evaluation, any new standard-setting methodology should be rigorously evaluated *before* making it an operational part of NAEP reporting. This should include an empirical evaluation of the achievement levels and an external validation of the results. Third, during the interim, current achievement levels should be accompanied by a warning stating that results should be interpreted as suggestive rather than definitive because they are based on a methodology that earlier evaluation panels have questioned in terms of accuracy and validity.



[29] See resolutions passed on May 13, 1992 and May 4, 1994 by the Assessment Subcommittee, Education Information Advisory Committee, Council of Chief State School Offices.

# 7 Reporting and Dissemination for the 1994 Reading Assessment

## Introduction

Previous chapters of this report have discussed the content validity of the 1994 reading assessment and the various technical procedures related to obtaining and analyzing TSA assessment data. This chapter addresses the final major component that must be considered in evaluating the *quality* and *utility* of the TSA program: the manner in which the data are reported and disseminated. Many of the Panel's guiding principles are relevant to this analysis, but particularly the *policy relevance* and *public information principles*. The first of these calls for timely reporting of data relevant to policy makers and education decision making, and the second for reports that are accurate in content, comprehensive in format, and readily accessible to all relevant stakeholders. For the TSA, state-level policy makers and educators are particularly important stakeholders, since these individuals have—not surprisingly—proven to be the primary consumers of state NAEP data. The extent to which the 1994 reading reports and related dissemination activities fulfilled the needs of this intended audience is therefore a key question.

Unfortunately, a full analysis of the 1994 reporting cycle is hampered by timing. Initial summary results for the national and state reading assessments were issued in a new, more concise format in April, 1995. However, there were a number of unanticipated delays both before and after the release of this *First Look* report, with the result that neither the individual state reports nor any of the other more comprehensive reports and data documents related to the 1994 TSA were released until March, 1996. The Panel therefore used what was learned from previous TSA reporting cycles to structure an evaluation based largely on dissemination plans, reviews of prerelease copies of the reports, and consideration of the factors impinging upon the report development process. The broader impact of reporting over all three TSAs is discussed in the final chapter, which summarizes the Panel's overall evaluation.

The current chapter is organized around four criteria that the Panel deems fundamental to successful reporting. In order of importance these are

- The accuracy of the results;
- The likelihood that the results will be interpreted correctly by the intended audience;
- The timeliness with which the results are made available; and
- The extent to which the results are accessible and adequately disseminated.

Throughout the chapter we have incorporated quotations from newspaper stories that followed the release of the April 1995 *First Look* report. These quotations illustrate the diversity of ideas about states' student achievement that were generated in response to the reported results.

With the advent of the TSA, a number of reporting issues emerged or became more prominent. For example, one set of technical issues arose as a result of the decision to produce separate customized reports for each state, and the need to coordinate reporting activities with the states generated various logistical concerns. In addition, although timeliness of reporting had always been a goal for NAEP, state-level reporting greatly increased the demand for more rapid access to results.

In 1990, the entire reporting process was revamped to provide individualized reports for each state as well as state-by-state comparisons. The NAEP contractor created a computerized reporting system to assist in this process, and NCES became more highly involved in all aspects of the reporting and dissemination activities. Procedures were also developed to obtain state input on proposed report formats and to allow states to review their own results before officially authorizing their inclusion in NCES reports.

For the 1992 TSA, these same reporting activities were continued, but with expanded effort because of the increased number of subjects and grades and the addition of short-term trend results in eighth-grade mathematics. Furthermore, "focused" reports that highlighted particular themes in the results (e.g., mathematics problem solving) were introduced for the first time in the 1992 cycle.

The 1994 report series was, in some ways, a reduced effort because the reports had to cover only a single subject and grade—fourth-grade reading—that also had been covered in 1992. The latter meant that some of the page layouts and explanatory materials could be reused in 1994. However, a number of innovations were also introduced, including new report formats, a separate preliminary release of summary results, and the first reports of state-level comparisons between public and private schools.

For the 1994 reading assessment, the following reports were developed:

- **The 1994 *First Look* reading report.** This relatively brief report (60 pages) was designed to convey essential information on national and state reading results in a user-friendly format with more "white space" and more graphics than earlier report formats. Intended to accelerate the release of the major results, the *First Look* report included only data that could be analyzed and summarized on a rapid turn-around schedule.

- **The 1994 *Reading Report Card.*** Following the *First Look* report, NCES released the more comprehensive *Reading Report Card for the Nation and the States* (180 pages). This report, in keeping with past NAEP report cards, contained detailed achievement results for the nation and the states and included findings from the student, teacher, and school questionnaires. In contrast to previous report cards however, the 1994 version was changed in various ways (e.g., by increasing the use of graphics and figures) with the intention of assisting the reader's interpretation of the data.

- **The state reading reports.** These customized reports (approximately 200 pages), each of which focused on the results for a single state, remained essentially the same as those produced in previous TSA reporting cycles and covered most of the same information as the *Reading Report Card*.

- **The 1994 data compendia.** The national and cross-state compendia provide student achievement results cross tabulated by each background variable derived from the student, teacher, and school questionnaires. The cross-state data compendium has been released in paper and electronic (CD-ROM) formats; the national compendium is expected to be available sometime in 1996.

- **Focused reports.** In addition to the reports listed above, plans were made to release focused reports highlighting NAEP findings of particular instructional or policy interest. These theme-oriented reports (none of which have yet been issued for 1994) are intended to provide extended coverage for results that might otherwise be buried in the masses of data generated by every NAEP assessment.

## *Accuracy of the Assessment Results*

One of the most crucial goals in reporting assessment results is to ensure that the data are accurate. This means that reporting schedules must accommodate thorough review and evaluation during the data analysis phase in addition to careful error checking of the reports themselves. Accuracy concerns are extremely important because the integrity of the results, and hence the quality of the information provided by the program, depend on how well this goal is met. In the history of NAEP reporting, great care has been associated with the reporting of the results and, for the most part, the data have not been questioned nor the results found incorrect. However, as discussed in chapter 5, the system is showing increasing evidence of strain under the increased burden of high volume, frequent innovations, and increased pressure for more rapid reporting.

> "In fourth grade, the only grade for which state-by-state scores were released, Georgia's girls' reading ability went down a little and boys' reading ability went down a lot. Georgia ranked about 39th among the 50 states. State school Superintendent Linda Schrenko labeled those results 'disappointing and unacceptable.' "
>
> *The Atlanta Constitution,* April 28, 1995

In particular, chapter 5 reviewed two events that occurred during the analysis of the 1994 reading results that carried implications for both the accuracy and the timeliness of the reports. In the first instance, unexpected declines in performance at grade 12 necessitated considerable additional work in an effort to confirm the validity of the short-term trend results. In the second instance, nearly all of the results had to be recalculated, and the *First Look* report had to be reissued, after two unrelated technical errors were discovered in the analyses.

To their credit, all parties acted promptly and professionally in explaining the problems to the public and taking corrective action where necessary. *The Panel commends NCES and the NAEP contractors for placing accuracy concerns foremost in the reporting decisions for 1994.*

## *Likelihood that Results Will Be Interpreted Correctly by the Intended Audience*

Following the release of the 1992 TSA results, the NAE Panel, NCES, NAGB, the states, and others re-examined the reporting process and recommended several improvements to enhance the reports and address various concerns. Issues included the best methods for presenting the data in the reports and the possible development of new types of reports to help focus the results for specific audiences.

Among the contributions to these deliberations was a report commissioned by NAGB that made a number of suggestions for improving the reporting and dissemination of NAEP data. Overall, the report proposed that NAEP publications be made more relevant to the classroom and instruction, be developed for more specific audiences, and be shorter and less technical. NCES and the NAEP contractor also met with a group of graphics experts and a team of external consultants to seek ways for improving the appearance of graphs and charts used in the reports. Additionally, the NAE Panel included several suggestions related to the interpretability of NAEP reports in its evaluations of the 1990 and 1992 TSAs. The overall goal of each of these efforts was to design reports that would be perceived as useful and intelligible by user groups and that would facilitate correct interpretation of the data.

> "Rhode Island's fourth-graders are winning new respect, and giving their elders a lot more satisfaction, when it comes to reading...The state Department of Education has announced that pupils improved their scores significantly, compared to their counterparts around the region and the nation, in the National Assessment of Educational Progress tests administered in February, 1994...The sharp improvement was credited by state education officials to the Literacy and Dropout Prevention Act adopted in 1988."
>
> *Providence Journal-Bulletin,*
> April 29, 1995

## *Conveying Statistical Significance*

One of the most difficult concepts to convey involves the statistical significance or insignificance of differences between states and across time. Some of the most innovative graphic formats included in TSA reports have been intended to address this problem. For example, each report since 1990 has included a chart, similar to those showing mileage distances between cities, that summarized the statistical significance

or insignificance of differences between each pair of states. The chart allows the reader to identify, for a given state, all of the states that performed significantly better than, as well as, or significantly worse than the target state.

Unfortunately, although the chart has worked well as a summary of between-state differences for readers with a reasonable degree of statistical sophistication, it has not been particularly well received by lay audiences or the press. The latter have tended to view the chart as too complex.

> "The [NAEP] figures released Thursday reflect a 12% increase since 1992 in the number of black fourth-graders reading below the basic level—defined as the skills needed to do schoolwork. The number of Hispanics below the basic level grew by 8%. 'I see that as a threat to our community at the same level as violence,' Chet Whye of the Rainbow Coalition said. 'Instead of a safe-city summit, maybe we need a smart-city summit.' "
>
> *Rocky Mountain News,*
> April 28, 1995

The 1992 and 1994 state reports included another type of graphic that addressed the same problem but simplified the presentation by showing the comparisons for only one state at a time. Specifically, a map of the United States was coded to show which states performed significantly better than, as well as, or significantly worse than the target state (in addition to identifying the states that had not participated in the TSA and were therefore not part of the comparison group). The resultant pattern was very easy to grasp visually and also worked well in a color-printed version that was provided separately and widely used in oral presentations of the results.

A variant of this map, coded to show short-term trend results (i.e., which states had significantly improved their performance, which had stayed the same, and which had declined significantly between assessments) was included in the 1992 *Mathematics Report Card*. This graphic was also very well received by the states, but, for reasons unknown to the Panel, was not included in the 1994 *Reading Report Card*.

## *Other Efforts to Improve Interpretability of Results*

Although the 1994 reports did not introduce any new solutions to the particular puzzle of conveying statistical significance, several format modifications were introduced into the *First Look* report and the *Reading Report Card* that were intended to generally facilitate the interpretation of the data. These included the use of more charts (including 3-D style bar charts), visual simplification of tables, more white space, and generally shorter reports. In addition, a glossy four-page summary was made available to facilitate wider dissemination of results, and a poster-size chart was developed to present a cross-state summary of the achievement results and selected contextual variables.

One might question the appropriateness of some of the reporting decisions, such as the very limited use of standard errors (or any other presentation of measurement error associated with the achievement estimates) in the *First Look* report. In general

however, *the Panel supports the improvements made to the 1994 NAEP reports and commends NCES and NAGB for their efforts to continually improve the interpretability and accessibility of their reports.*

## *Timeliness of Reports*

Table 7.1 shows the time to reporting for each of the TSAs. Although shorter turn-around time has been identified as a priority goal since the inception of the state NAEP assessments, one can see that success has been, at best, quite mixed. In 1990, 16 months elapsed between the administration of the TSA and the release of the main state and national mathematics reports. Although this was slower than desired, it was considered reasonably acceptable as a first effort in view of the fact that so much new machinery had to be put in place. However, hopes were high that the time lines could be significantly shortened for 1992.

**Table 7.1.** ***NAEP TSA reporting time lines, 1990-1994***

| Assessment Year | TSA Subject Area | Release Date |
|---|---|---|
| 1990 | Mathematics—main reports | June 1991 (16 months)* |
| 1992 | Mathematics—summary findings (national only)<br>Mathematics—main reports<br>Reading—main reports | January 1993<br>April 1993 (13 months)<br>September 1993 (18 months) |
| 1994 | Reading—summary findings (national and state)<br>Reading—main reports | April 1995 (13 months)<br>March 1996 (23 months) |

* The release of the 1990 TSA results was 16 months after the completion of the TSA administration, which ended in March, but 13 months after the completion of data collection activities for the main national assessment, which ended in May. In subsequent years, most of the data collection activities for the main assessments were completed on the same schedule as the TSA.

In fact, a number of factors, including various issues related to reporting by achievement levels, delayed matters so that, in the end, time to reporting was roughly equivalent to 1990. Although a highly abridged summary of the 1992 national mathematics results was released after only 10 months, no TSA results were available until the main mathematics reports were released in April, 1993 (13 months); state and national reading results followed in September, 1993 (18 months). NABG and other TSA constituencies were less tolerant of these extended time lines for the second TSA, and timeliness of reporting again became a major point of discussion in planning for 1994.

Once again substantial improvements seemed possible, particularly in view of the fact that the NAEP contractors continued to make great strides in computerizing many of the time consuming activities that occurred after an administration (e.g., scoring the constructed-response items via an image processing system and creating a computer-generated reporting system that automatically customized the state reports). Specifically, a schedule was established that called for the new *First Look* report, which would contain the central findings for the nation and the states, to be released within the same calendar year as the assessment (i.e., by the end of 1994), and for the more comprehensive state reports and *Reading Report Card* to be available by September, 1995.

Unfortunately, although the Governing Board, NCES, and the contractors all made intensive efforts to meet the announced schedule, this did not occur. Rather, as summarized in table 7.1 and shown in greater detail in table 7.2, the *First Look* report was not released until April, 1995 (13 months), and, although no one would have expected it at the time, the main reading reports did not appear for nearly another year after that. *This two-year lag between the assessment and the release of the 1994 Report Card was the longest in the six-year history of TSA reporting.*

What went wrong?

## *Analysis Problems and Competition for Resources*

First, it must be understood that data processing, scoring, scaling, and analyses account for most of the time between the administration and the release of reports. Activities narrowly related to the production and approval of report documents also take time, but not as much as these other preparations. Of all the steps involved, scaling and analyses present the greatest uncertainties—particularly when new content frameworks are being introduced or new types of analyses undertaken—and therefore the greatest potential for unanticipated delays and schedule slippages. Furthermore, the complexities of the NAEP design appear to have created a situation in which the number of individuals competent to oversee and trouble shoot the analyses, even among the contractor staff, is extremely limited. Consequently, when many nonroutine activities are underway simultaneously, there is significant competition for human resources.

In the case of the 1994 reading report schedule, original plans had called for the NAEP contractor to put all or most of its resources toward the rapid completion of the scaling and analysis work to support the reading reports. However, competing requests were soon made by NAGB and its achievement-levels contractor, who needed item performance data in order to proceed with standard setting for the other subjects assessed in 1994 at the national level: U.S. history and world geography. This caused a reallocation of staff commitments at ETS and a rescheduling of plans for the release of the reading data (and was one of the initial reasons that the national and state reading reports were not released as originally scheduled). Next, prior to the release of the *First Look* report in April, 1995, extensive, unanticipated effort was required to arrive at an appropriate set of common items to use in establishing the short-term trend results and to investigate the possibility that the grade-12 score decline might be artifactual rather than real. All of this again drew off resources and caused further delays. Third, the discovery of technical errors by both ETS and ACT

**Table 7.2.** ***Reports and release dates from the 1994 NAEP TSA***

| Report | Originally Planned Release Date | Actual Release Date |
|---|---|---|
| *NAEP 1994 Reading: A First Look*—summary and highlights of national results for grades 4, 8, and 12 and state results for grade 4 | December 1994 | April 1995<br>October 1995 (rev) |
| *NAEP 1994 Reading Report Card for the Nation and the States*—full national and state results | September 1995 | March 1996 |
| State reports—customized reports for each state | September 1995 | March 1996 |
| "Major Findings for the Nation, Regions, and States," Executive Summary of the *NAEP 1994 Reading Report Card for the Nation and the States*—four-page highlights of full report | September 1995 | March 1996 |
| *Technical Report of the NAEP 1994 TSA Program in Reading*—details of design, operational, and analytical activities | September 1995 | March 1996 |
| *Cross-state Data Compendium for the NAEP 1994 Grade 4 Reading Assessment*—detailed tables of state data | September 1995 | March 1996 |
| State Restricted Use Data Tapes | late 1995 | * |
| National Reading Data Compendium—detailed tables of national data | mid 1996 | * |
| Further Exploration of Reading Results—focused report | late 1995 | * |
| NAEP 1994 Reader (grades 8 and 12 national data)—focused report on a special study that looked at the impact on assessment results when students were allowed to choose among alternative reading passages. | early-mid 1996 | * |

* Not yet released

after the April *First Look* release necessitated new data analyses, a revised *First Look* report, and another significant delay. Fourth, because of the delays already encountered, NAGB decided that the U.S. history and world geography *First Look* reports should be given priority so that results for these subjects could be included in the Goals Panel annual report, even though this would cause further delays for the remaining reading reports. Finally, as an additional irony, two government shutdowns and a blizzard sabotaged plans to release the main reading reports in December, 1995. The reports were finally released on March 7, 1996, 23 months after the assessment.

## *State Review of Results*

The authorizing legislation for the NAEP state assessments requires that each state be given the opportunity to review its own results before approving their release. Therefore, ETS must prepare and disseminate preliminary state results, and build in time for the states to examine their results and respond with their reporting decisions. Although report preparations continue during this period, additional delays can be encountered if a state declines release (as occurred in both the 1992 and 1994 TSAs). In the latter case, data for the declining state must be expeditiously removed from the reports and significance tests recomputed for many of the cross-state and cross-time comparisons.[1]

***Although state review is just one of many factors adding time and uncertainty to the release schedule, the Panel generally questions the wisdom of allowing states to withdraw once the data have been collected and have met the minimal technical criteria for reporting. For this reason, the Panel suggests that Congress, with input from the states, reconsider this proviso when NAEP is reauthorized in 1998.***

## *NCES Adjudication Process*

The adjudication process is another component in the time line to reporting. NCES requires that every report it publishes be adjudicated by technical experts to make certain that it meets the agency's statistical and reporting standards. This is an extremely valuable quality control process that unquestionably should be maintained. However, observers both inside and outside of NCES have noted that the process, particularly as it pertains to NAEP reports, is not optimized for speed. Currently, the adjudication process adds at least one or two months to the overall time line for reporting. Delays can be substantially increased if adjudicators identify methodological concerns or patterns of findings for which they believe additional confirmatory analyses should be performed (and the new results subsequently reviewed).

[1] A second and much less problematic review of state results occurs just before the NCES release date to allow states to review the results of the state comparisons and prepare for questions from their constituents and the press. This latter activity adds only slightly to the time to release.

Clearly, problems are worsened when NAGB or NCES change the analysis or reporting requirements midstream, when novel analyses are pursued on tight time lines, or when coordination between contractors, project officers, and adjudicators is not carefully planned and maintained.

### *Summary*

Because NAEP is a key indicator of students' educational progress, it is important that the nation and the states be able to count on the timely release of NAEP assessment results. For this reason, the Panel strongly encourages NCES and NAGB to continue to press for quicker and more timely reporting while also being careful to maintain the quality and integrity of the data. To NCES's credit, the accuracy of the results took precedence over timely reporting in the case of the 1994 reading assessment. However, the serious delays in releasing the main reading reports almost certainly lessened the impact of the results. The Panel recommends that NCES and NAGB continue to explore ways to expedite all aspects of the process, including analysis, quality control checks by the contractor, report preparation, and NCES adjudication.

Obviously, time from collection of the data to its release is not independent of the overall NAEP design. Because of this relationship, the Panel will also address the issue of timely reporting in the larger context of its upcoming capstone report.

## *Dissemination and Accessibility of the Findings*

### *Release and Press Coverage of the **First Look** Report*

Following the pattern set by the first two TSAs, a national press conference was held by NCES and NAGB to announce the release of the *First Look* report on April 27, 1995. Individual state results were embargoed until the national release, after which states were free to publicize their own data; in many cases, the states simply chose to piggy-back on the publicity generated by the national press conference. In general, the 1994 release, like its predecessors, proceeded smoothly. The one problem noted by some states was that copies of remarks made at the national press conference by top education officials were not available to the states in advance. Because these remarks generally determine the "spin" that the press will put on the results, state officials were not as well prepared as they might have been to respond to questions.

Logistically, NCES provides the reports and related publicity materials to the state departments of education, and they, in turn, further distribute the data and present the information to the districts and schools in their states. This "trickle down" approach has worked very well in some states, but not as well in others. NCES also uses mass mailings to distribute some reports directly to local districts and schools. NAEP has not, however, established much name recognition at the local level, and to date no systematic information has been collected regarding the impact of these "cold mailings."

In order to monitor press coverage of the *First Look* report, the Panel undertook a review of NAEP-related articles appearing in the 50 most read newspapers in the United States during the weeks following the April 27 release. Similar efforts were carried out after each of the previous release dates for TSA reports. In general, the Panel found that the volume of coverage decreased with each cycle of the TSA. For the 1990 reports, coverage was brief (generally only one article), but extremely widespread. Less coverage was observed for the 1992 reports, and still less for the 1994 *First Look*. Although the overall trend was down, press coverage was more extensive in some states, as the quotations inserted throughout this chapter suggest.

> "[T]he U.S. Department of Education released results from the National Assessment of Educational Progress that showed little change from 1992 to 1994 in Kentucky's fourth-grade ability to read. These results disturb some because the states' own test, created as part of the 1990 Kentucky Education Reform Act, showed students improving dramatically in the same subjects at roughly the same time...Some critics of school reform claim the latest results show that Kentucky's own test is fatally flawed, and even people who are more supportive say the results could indicate problems."
>
> *The Courier-Journal,* May 10, 1995

The majority of the 11 states that evidenced the most press coverage for the *First Look* report had performed poorly relative to other states or relative to their own performance in 1992, and the news articles correspondingly emphasized the poor performance of their students. Only two states, Nebraska and Wyoming, reported their results in a positive light. These two states had average proficiency scores that were among the highest in the nation. In California, which was tied with Louisiana for the worst average achievement scores among the states, an especially large amount of press coverage was noticed. Depending upon the individual perspectives of their sources, the articles tended to explain the state's poor showing by blaming different factors. Some blamed the whole language reading approach, others blamed the large proportions of LEP and transient students, too much television, too little spending on education, or overcrowded classes. Some of these explanations were not unique to California; similar statements were also made in other states not performing up to expectations.

Other news articles contrasted the decline in national and state NAEP reading scores with the (1990-1992) rise in mathematics scores, or related the decline in reading scores to various phenomenon such as the concomitant rise in time spent watching television (as reported on the NAEP student questionnaires) or the lack of parental involvement in their children's education—particularly parents who were not reading enough to their younger children or who were not sufficiently involved in their children's homework. None of the latter factors are measured by the NAEP variables.

In keeping with all of NAEP reports, the 1994 reading reports warned readers against making invalid conclusions, especially those that imply causation. However, as can be seen from the news reporting, such conclusions continue to be drawn. The Panel encourages NCES and NAGB to continue efforts to remind users of the data, and particularly the press, that it is not valid to attribute student performance on NAEP to

other factors, whether or not the latter are measured by NAEP. It is important to note that the assessment data can only show statistical relationships among the variables and not cause-and-effect relationships.

## *Release of the* ***Reading Report Card***

The full set of 1994 reading assessment reports, which included the *Reading Report Card,* a four-page summary brochure, the customized state reports, a cross-state data compendium, and the technical report for the 1994 TSA, were finally released on March 7, 1996, 11 months after the original *First Look* release and nearly two years after the data were collected. The reports were released without either a press conference or a formal press release and, not surprisingly, were treated as a virtual nonevent by the press. It appears as if only the education press gave the reports any space at all.

> "Although Maine fourth-graders ranked first in the nation for their reading ability, only a third of those children were judged proficient readers, according to a national survey released Thursday. While educators praised Maine's performance on the National Assessment of Educational Progress, they admit there is still room for improvement."
>
> *Bangor Daily News,* April 28, 1995

Because this report on the Panel's evaluation was written shortly after the March release, there was no opportunity to evaluate the other ways in which the newly available TSA reading reports might have impact. For example, the state-level private school results for those states that met the requisite participation standards appeared for the first time in the *Reading Report Card.* These may eventually engender press coverage and debate despite the low-key nature of the release.[2] Moreover, the experience of the previous TSAs suggests that many of these newly released materials will be used by state assessment directors and reading specialists in the states that participated.

After previous assessments, for example, state education agency (SEA) staff in some states spent considerable time reviewing the detailed information provided by NAEP on teaching practices, and the relationships between these reported practices and student achievement. SEA staff have also been interested in achievement results for population subgroups in their own states, and in the sample assessment tasks and examples of student responses that have been included in NAEP reports. In some cases, these materials have also been used in in-service training programs or conferences given by SEA staff members. Finally, state assessment directors have previously expressed a need for brief and readable summaries of NAEP results that they could distribute to educators and policy makers in their states. The four-page brochure released with the *Reading Report Card* would appear to serve this function,

[2] As discussed in chapter 3, the Panel does not endorse the separate reporting of public and private school results at the state level, finding that there is considerable risk of misinterpretation, particularly in light of the small samples and nonrandom participation patterns for private schools.

although it may be of limited use in view of the amount of time that has passed since the publicity of the *First Look* report.

### *Other Reports*

NCES also plans to release two focused reports from the 1994 reading assessment at some point in 1996 or 1997. Such reports were recommended by the Panel and others after the first TSA, and several were introduced for 1992. A few of the latter received a fair amount of press coverage, and all appear to have been well received by subject area specialists. One can assume that the 1994 focused reports, when they appear, will be similarly well received.

## *Suggestions for Increasing the Accessibility of NAEP Data*

A substantial investment has been made by Congress and the states in state NAEP, and the program has produced extensive, high quality data. However, there is a persistent perception that these data are under utilized. The following paragraphs offer some suggestions that appear to have good potential for increasing accessibility.

### *More Involvement of the Press*

Further efforts should be made to identify ways of making NAEP information accessible and interesting to the press. Currently, many news stories merely pick up the language from the official press release. This is fine as far as it goes, but input from the media ought to be sought to identify strategies that could effectively encourage greater use of NAEP findings and provide guidance on how to develop accurate, useful, and interesting news stories from these data. For this reason, the Panel repeats, in slightly altered form, a recommendation made in its evaluation of the 1990 TSA, *Assessing Student Achievement in the States*.[3]

The Panel recommends that NCES continue to provide the media with specific examples of appropriate interpretations of the NAEP results. NCES should also continue efforts to construct new and more effective reporting formats that clarify the important relationships in the data and discourage inappropriate interpretations. Such efforts may be facilitated if the new formats are presented for comment in focus groups involving the media, prior to the release of future reports.

[3] See The National Academy of Education, *Assessing Student Achievement in the States* (Stanford, CA: Author, 1992), 52-3.

Feedback from user groups has suggested that the large, multipurpose report cards prepared in 1990 and 1992 were daunting to all but the most interested and informed consumers. The 1994 *Reading Report Card* was pared down considerably, but still came close to 200 pages. Furthermore, much interesting data was not reported in any source but the archival data compendium. A more appropriate strategy might be to plan the release around multiple reports that break down information into more manageable sections and target the needs and interests of specific audiences. For example, information on students' home support for literacy or use of instructional materials could be summarized in separate "mini-reports" suitable for parents or classroom teachers. Other, perhaps more technical, reports could provide education professionals with extended coverage of interesting topics such as students' abilities to adopt different reading stances. (The latter would be very much like the focused reports introduced for 1992, which NCES has already indicated its intention to continue.)

> "A survey that was part of the 1992 NAEP results—in which California fourth-graders also came in near the bottom—found that the state's teachers were far more likely than their counterparts in any other state to believe in the so-called language approach to reading."
>
> *Los Angeles Times,*
> April 28, 1995
>
> "California made a horrendous mistake in taking out the phonics and the basic decoding skills from our reading programs, and when you do that, kids aren't going to learn to read anywhere well enough, if at all."
>
> Maureen DiMarco
> Education Advisor
> to Governor
> Pete Wilson

The cross-sectional design of NAEP makes it impossible to draw causal inferences. The reports, however, could discuss the observed relationships between student achievement and such factors as teachers' reports of the frequency of particular classroom practices, so long as they clarified for the reader the kinds of inferences that could and could not be made from the available data. In many cases it might be possible for the reports to include corroborating evidence from other data sets, such as the National Education Longitudinal Study (NELS:88), that employ longitudinal or experimental designs. In other cases, the reports should clearly specify that any causal inferences about the observed relationships are only speculative, although they may prove useful for suggesting hypotheses for further study. NCES should draw upon the Department of Education's Office of Educational Research and Improvement (OERI) to help locate extant research findings that corroborate correlational findings from the NAEP data sets and can be used to enrich the focused reports.

NCES is to be commended for developing a series of focused reports that highlight specific findings from NAEP assessments. The Panel recommends that NCES continue and expand this approach, drawing on other research where possible to corroborate and inform the correlational relationships observed in the NAEP data and, at the same time, clearly warning readers against unwarranted causal inferences.

### *Provide More Examples of Assessment Tasks and Student Responses*

SEA staff have noted that certain members of the potential NAEP audience, especially teachers and parents, find the inclusion of concrete examples of assessment tasks and student responses extremely helpful as they examine even summary-level NAEP results. The Panel was disappointed not to see sample tasks included in the *First Look* reading report. Because such examples were included in the subsequent *First Look* reports for U.S. history and world geography however, it appears that this practice will be continued in the future. ***The Panel suggests that NCES include plentiful examples of assessment tasks and student responses in all NAEP reports so that readers can better understand the basis for the scaled results. If necessary, item development schedules should be adjusted to ensure that appropriate released items (i.e., items that are not needed for future assessments) will be available for this purpose.***

### *Involve the States More in Reporting and Dissemination*

When considering the various alternatives for reporting and dissemination, it is important to weigh and compare costs and benefits. Although it may be quite useful to provide additional types of reports and to disseminate them more widely, or to produce more focused reports for narrower audiences, each of these suggestions would probably increase overall costs. Alternatively, NCES and NAGB might consider *reducing* centralized reporting in favor of assisting the states in producing their own reports, customized to their own situations and needs.

The Panel was pleased to learn that just as this report was being completed, NCES had put a special NAEP home page on the World Wide Web (http://www.ed.gov/NCES/surveys/naep.html) in which selected results from the 1994 reading assessment are displayed. Such an effort is very much in the spirit of providing an infrastructure that the states (and others) could draw upon for creating their own customized reports, and the Panel commends NCES for taking this step toward innovative dissemination of NAEP results.

The Panel has not systematically examined the question of how heavily states would be willing to invest in producing their own NAEP reports, but if this could be done with limited effort using the types of menu driven software that have already been investigated by NCES and ETS, the alternative could prove palatable. Minimally, using the Internet to *distribute* reports prepared by NCES has the virtue of saving printing and distribution costs, and the Panel is aware that NCES has already taken steps in this direction.

> "South Carolina students ranked near the bottom in a national test of reading skills, as more than half failed to achieve even basic reading levels...State education superintendent Barbara Nielsen said that the scores come as no surprise. 'I know we need to raise our standards, and we will raise our scores,' she said. 'When children are expected to do more, they will do more.' "
>
> *The Post and Courier,* April 28, 1996

Another way of providing centralized assistance for state dissemination efforts would be to collaborate with states to offer workshops for district and school-level educators. Such workshops could provide hands-on opportunities to show local educators how to use NAEP materials and findings to inform classroom practice.

## *Summary*

In summary, the Panel concludes that the reporting of the 1994 TSA, though comprehensive and thorough, was also problematic in some respects. In particular, although a summary *First Look* report was issued in April, 1995, the main reading reports for the nation and the states were not released until March, 1996—fully two years after the assessment had concluded and, in fact, after most of the data collection for the 1996 state assessments had occurred as well! Furthermore, problems with the data analysis, discussed in chapter 5, not only contributed to the long delays in reporting, but also necessitated the release of a revised *First Look* report in October, 1995. On the positive side, the Panel notes that the reports prepared for 1994 included some attractive format innovations intended to make the data more accessible and easier to understand.

In considering the issues surrounding reporting and dissemination, it is important to keep in mind the Panel's principles of *quality, utility, public information,* and *policy relevance.* The Panel cannot over emphasize the importance of accurate reporting and interpretation of results. The press for accuracy, however, can clash with the press for speed, particularly when new content frameworks and analysis procedures are involved. Pressure to release reports quickly while maintaining the accuracy of the data was a theme in both the 1992 and 1994 NAEP assessments, and it is worth noting that the specific analysis errors that led to the reissuing of the 1994 *First Look* report actually occurred in 1992 and had simply been perpetuated. Until such time as NAEP adopts a simpler design better suited to the demands of the current situation, the tension between accuracy and timeliness is likely to remain.

Finally, it is important to continue efforts to enhance the accessibility and interpretability of NAEP reports and to suit them to their intended audiences. In the Panel's opinion, much work has already been accomplished toward meeting these goals, and some suggestions for further enhancements have been offered here.

# 8 Conclusions and Recommendations

## Introduction

In 1990, when the Panel began its evaluation of the first trial state assessment, 37 states, the District of Columbia, and two U.S. territories agreed to participate. The attitudes of the participating states was guarded, however, anticipating the possibility of negative as well as positive consequences. Some, for example, were leary of an overemphasis on the "horse race" created by the cross-state comparisons; others hoped to see the assessment boost public interest in education and accelerate the reform of classroom practices.

As the trials continued, a few of the originally participating jurisdictions dropped out, but the total number of participating states increased slightly with each trial. By the time of the 1994 TSA, 41 states participated,[1] and, having seen that the negative consequences they feared were generally not materializing, states' attitudes had become more uniformly positive. Although the TSA's positive impacts fell short of the very high expectations that some had held, the assessments did provide important opportunities to measure student achievement against a common set of challenging, consensus-based standards, to validate or supplement information from their own state assessments, and to track the progress of their students over time and in comparison to students in other states. Without question, by 1994, state NAEP had become an important and respected part of the assessment landscape.

This report summarizes the Panel's evaluation of the 1994 TSA in reading. The conclusions and recommendations, like those in the Panel's previous reports, are strongly rooted in empirical research commissioned by the Panel and in the Panel's guiding principles.

The first and most essential purpose of this report is to provide information that Congress can use as it deliberates levels of funding for NAEP in upcoming years and the reauthorization of NAEP in 1998. A second important purpose for the Panel's suggestions and recommendations is to provide information that the Governing Board may consider as it sets policy for state NAEP. The third and final purpose is to provide guidance to NCES and its contractors in resolving questions related to the conduct of state NAEP.

## The Success of the 1994 TSA

As noted, the current report focuses on the 1994 reading assessment and, in particular, on the state assessment in reading at grade four. As it did in its previous reports, the

[1] Guam and Department of Defense overseas schools also participated in the 1994 trial, and the District of Columbia participated, but withdrew after the data collection phase. Public school results for two of the 41 participating states were not reported because they failed to meet minimum school participation guidelines.

Panel addresses the major topics of assessment content, conduct, and reporting. In addition, some new topics have been given special emphasis for 1994, including the assessment of IEP and LEP students, the continued use of the reading achievement levels for reporting, and the adequacy of the current procedures for analysis and scaling. In each of these areas of inquiry, the Panel's findings have been generally very positive, attesting to the high quality of the NAEP program and the state NAEP assessments. At the same time, the Panel also has identified a number of factors that should receive further attention due to evidence of system strain, anticipated growth, or likely benefits from new research. The following sections briefly summarize the Panel's major findings for each topic.

## *Content Validity*

The 1994 NAEP reading assessment marked the second use of the reading framework developed in 1991. A portion of the item pool was released to the public and replaced between the 1992 and 1994 assessments, but the overall parameters of the assessment were held constant, allowing reading trends to be measured for the first time on tasks that reflect current understandings of reading and reading assessment.[2] *The Panel reviewed the framework and items for content validity after each of the two assessments and, in each instance, concluded that the NAEP reading assessment was a reasonable representation of current theories in reading, a valid measure of reading achievement in the nation, and was relevant to everyday classroom practice.*

Although the Panel's reading experts also noted aspects of the framework and item pool that could be improved, none of these shortcomings were sufficient to undermine the content validity of the 1994 assessment. Furthermore, the Panel concluded that the decision to hold frameworks in reading and other content areas constant over several assessment cycles was praiseworthy—a judgment that was confirmed by the strong interest of NAEP constituents in using 1994 results to gauge the progress of their students over time.

More specifically, the Panel commends NAGB and NCES for building a challenging assessment of reading achievement that extends beyond simple mastery of the mechanics of reading to include the reader's ability to draw meaning from text and to communicate this understanding to others. At the same time, the Panel notes that the 1994 fourth-grade assessment contained relatively few items that were within the scope of the least able students, making it difficult to get precise and reliable estimates of achievement for those at the lower end of the scale. Some unevenness of item quality was also observed. Specifically, some of the scoring guides for constructed-response items were inconsistent with other features of the items or with the directions given to students, and a number of the more difficult items failed to capture the essential features of advanced reading achievement. The Panel judged all of these to be areas in which improvement should begin immediately and was pleased to learn that the NAEP contractor had already taken steps to improve the scoring guides and clarify directions to students.

[2] Previously NAEP had relied only on the much older instruments from the long-term trend assessments to measure trend; the program now has two overlapping trend measures that can be linked together before the older measure is eventually phased out.

Other shortcomings were in areas that, for one reason or another, the Panel believes should be left untouched for now. Included in this group is the omission of items measuring reading to perform a task at grade four, a framework decision that gives the unfortunate impression that reading to perform a task is not within the scope of fourth-grade students. However, although this type of item could be developed with relatively little difficulty and added to the fourth-grade assessment, any such change should be left for the next framework revision in order to protect the measurement of trends.

By contrast, the Panel judged that more research will be needed in order to clarify reading stance—one of the two main organizing dimensions of the framework—or to systematically address the impact on reading of prior knowledge, personal background, and experience. When problems like the latter are solved, the ability of the framework to guide the development of assessment tasks and facilitate the interpretation of reading achievement results will be enhanced. In order to make substantive progress in these and other areas, the Panel recommends that NAEP development cycles be modified to allow ongoing research and development, and to permit innovative concepts and item types to be thoroughly pretested and refined on the basis of empirical results before new frameworks are used operationally.

## *Sampling and Assessment Administration*

The sampling and administration procedures used in the 1994 TSA closely paralleled those used in the two previous TSAs. *As it had in 1990 and 1992, the Panel concluded that both sampling and administration for the 1994 TSA were done well and were generally consistent with best practice for major surveys of this kind.* Two areas of concern were identified however.

First, and most importantly, substantial problems were found with the samples of nonpublic schools that were—at the Panel's previous recommendation—added to the TSA for the first time in 1994. The samples, which were originally intended to support only composite reporting of public plus nonpublic school results, were too small to serve as the basis for the separate reporting of nonpublic school results that ultimately occurred. Furthermore, participation rates for the originally-sampled schools were unacceptably low in approximately 40 percent of the states.

The Panel's motivation to include nonpublic schools was based on its *inclusiveness principle,* and the Panel's intention was to aggregate the nonpublic school results with those from public schools in order to generate overall state composites. In view of the fact that states compare themselves with each other and that the percentages of nonpublic school students vary widely from state to state, the Panel thought it important to report TSA results for the public/nonpublic composite, along with results for the public schools alone. However, NCES found it difficult to recruit nonpublic schools without offering them separate reports of student achievement by type of school. Even with the separate reporting incentive, the participation rates of nonpublic schools varied considerably from state to state and were unacceptably low in many. *The Panel recommends that nonpublic school results* ***not be reported separately*** *because of the considerable potential for drawing incorrect inferences about differences in performance between public and nonpublic schools.* Furthermore, it is the Panel's understanding that nonpublic schools do not have a strong interest in

comparing themselves with public schools at the state level. They are, however, very interested in comparisons at the national level.

For this reason, the Panel believes that separate nonpublic school reporting at the state level is unnecessary. Instead, the Panel urges that the nonpublic school data collected at the state level be combined with nonpublic school data collected at the national level and that results for this larger sample be reported at the national level, categorized by type of nonpublic school. Meanwhile, the results at the state level should be reported only for all students combined and for public school students alone. Furthermore, the reports should include prominent warnings about the factors that make public/nonpublic school comparisons based on NAEP results inappropriate, whether at the state or national level. Finally, the Panel's strong concerns notwithstanding, if NCES persists in reporting the nonpublic school results at the state level, it should expand the size of the nonpublic school samples and develop effective incentives for participation by nonpublic schools in order to produce results that will support more trustworthy inferences.

A second area of concern involves the participation of *public* schools. Although the Panel found that, for nearly all states, the participation rates for originally-sampled public schools ranged from acceptable to good, strong indications have emerged that the burden on the states, especially small states, may begin to threaten school and hence state participation rates in years when multiple subjects and grades are assessed. As a result, the Panel believes that NCES should consider design changes that could decrease the burden on all, but especially the smaller states, without compromising the overall quality of the assessment.[3]

## *The Assessment of Students with Disabilities or Limited English Proficiency*

The 1994 assessment cycle occurred at a time when NAEP was beginning to re-examine its policies regarding the exclusion and assessment of students with disabilities or limited English proficiency. In 1994, NCES gathered additional data from several sources to help its deliberations: 1) questionnaires covering level of functioning and education experience were collected for the first time for included, as well as excluded, IEP and LEP students; 2) the Panel, as part of its evaluation, collected new data for samples of IEP and LEP students who had been selected for participation in the TSA; and 3) NCES held conferences with representatives of the disability and bilingual communities to discuss the best methods for increasing inclusion in NAEP. *The results of these various efforts were laudable, leading to a set of revised exclusion procedures and new allowances for accommodated assessment that were tried out in the 1995 field test and implemented, in a controlled design, in 1996. The 1996 design is particularly commendable due to the fact that it allows NAEP to move forward with its important inclusion agenda while still ensuring that trend lines*

[3] One suggestion is that NCES explore the use of finite population sampling techniques, which are more appropriate to small populations, with the aim of possibly reducing the burden on small states.

*can be maintained and that the impact of the inclusion and accommodation changes on NAEP results can be investigated.*

The results of the Panel's 1994 study showed that changes were needed with regard to school personnel who, in different states, tended to interpret the (old) exclusion guidelines differently. Thus, on average, IEP students with the same level of ability would be included in some states and excluded in others. The Panel also found that a high proportion of IEP students (perhaps as many as 85 percent) could read well enough to participate in NAEP and be included in estimates of overall state achievement,[4] and that teachers of both IEP and LEP students were likely to propose testing accommodations for high percentages of their students. An implication of the latter finding, borne out by the experience of the 1995 field test, was that, if accommodations were offered, inclusion would be increased, but the overall numbers of students assessed under standard conditions would actually go down. This is problematic because scores attained under nonstandard conditions are much more difficult to interpret.

## *Scaling and Analysis*

The procedures used for scaling and analysis in the TSA are generally the same as those used in the national NAEP. However, state NAEP complicates analyses overall in that many steps must be performed separately for each state, linkages between state and national populations must be established, and appropriate between-state comparisons must be computed. In 1994, two technical errors affecting state scores were discovered in different parts of the analysis, and an unexplained but statistically significant drop in performance (which may have been, at least in part, an artifact of the assessment and analysis procedures) was observed in the national reading results at grade 12. These factors led the Panel to give greater attention to scaling and analysis in this evaluation than it had in its previous evaluations.

While examining factors that might have contributed to the decline in 12th-grade reading scores, the Panel determined that the NAEP contractor had had some problems linking the 1992 and 1994 reading assessments. Because the statistical characteristics of many of the constructed-response items did not remain constant across the two assessments, these items had to be eliminated from the common pool of items that formed the basis for the cross-year equating link. The common item pool consequently had too high a proportion of multiple-choice items. If the two item types indeed measure somewhat different skills, then the link was not a good proxy for the whole.

Furthermore, because no standard errors for the NAEP equating are computed, it is not possible at present to obtain estimates of the amount of additional error (i.e., uncertainty) due to equating that should be added to the total error for the trend

[4] The Panel's study of LEP students was more limited and did not allow parallel conclusions to be drawn, although there were indications that a significant proportion of LEP students also read well enough in English to participate for the purpose of contributing to overall state NAEP results.

estimates.[5] For this reason, the Panel recommends that a study to determine equating error be undertaken by NCES or its contractor. The inclusion of the standard error of equating as a component of the overall standard error of year-to-year performance estimates would provide an improved basis for evaluating change.

The study of the decline in 12th-grade reading scores also alerted the Panel to the need to have other data results easily accessible to provide some confirmation for the new short-term trends. The Panel noted, for example, that the long-term reading trend data for 1994 had not been analyzed on a schedule that made them readily available for comparison with the short-term results. Nor had other data sources that might have helped to corroborate (or failed to corroborate) the drop in 12th-grade scores been compiled and made available (e.g., trend results in reading from various state assessments). The Panel suggests that in the years when long-term NAEP trend data are collected, these data should routinely be available for comparison with the short-term trend results. The Panel further suggests that any significant change in performance on the short-term trends routinely be checked against other sources of trend data—sources such as the long-term NAEP trend data and state assessment trend data—before the results of the short-term trend are reported.

The Panel concludes that NCES and its contractors continue to make use of sophisticated methods to solve challenging measurement problems posed by recent innovations in testing (e.g., accommodating partial-credit items in the scaling model) and to produce generally high quality data. At the same time however, the system appears to be showing strains that allow errors to creep in, in addition to lengthening the time to reporting. Factors contributing to these strains are the pressure of rapid change (new enhancements have been added at every assessment that then must be accommodated statistically in the analyses), increased volume, and policy pressure to reduce time to reporting. More fundamentally, the underlying NAEP design, though well suited to NAEP as it was carried out in the mid-1980s when the design was adopted, may no longer fit the size and objectives of the current NAEP program. For this reason, the Panel supports NAGB's efforts to develop a new, more streamlined design for NAEP. In the meantime, quality control measures should be strengthened as much as possible, and the contractor should continue to carefully document all procedures in technical reports that are available concurrently with the main results.

Finally, on a separate note, the Panel notes that, *in view of the large number of comparisons (e.g., between states and between years) that must be computed for state NAEP, the traditional Bonferroni correction is overly conservative and fails to detect large numbers of "real" differences.* More powerful procedures are available, and the Panel recommends that they be explored.

## *Achievement Levels*

The achievement levels established by NAGB for the 1992 reading assessment were again used for reporting the 1994 assessment. From one perspective, the Panel understands NAGB's decision in view of the fact that reporting by performance

[5] For every score estimate reported by NAEP, the standard error indicates the width of a band around that estimate in which the true score might reasonably lie. The larger the standard error, the lower the certainty that the reported score is very close to the true score.

standards is greatly valued by much of the NAEP (and TSA) constituency. Nevertheless, the Panel points out that its extensive evaluation of these achievement levels, conducted in 1992, raised serious questions about their reliability and validity. At that time, the Panel's review suggested that 1) the standard-setting method had led to serious internal inconsistencies that could have especially troubling consequences if the mix of item types changed over time, and 2) the distributions of student performance established by the achievement-level cutscores was not reasonable based on comparison to the distributions suggested by various non-NAEP measures. In particular, the weight of the evidence suggested that the 1992 achievement levels were set too high.

Although the achievement-levels contractor fielded a study in 1994 that putatively addressed the second of these concerns, the Panel concluded that the design of the study did not permit confirmation of specific cutscores. The study was therefore not particularly informative with respect to the Panel's conclusion that the cutscores had probably been set too high. With regard to the Panel's other concern about fundamental flaws in the standard-setting method itself, NAGB and NCES did jointly sponsor a conference on standard setting that presented some interesting ideas for other ways in which performance levels might be set. However, NAGB has not, to the Panel's knowledge, sponsored any research to test the viability of other methods.

The Panel also examined the results for the 1994 U.S. history and world geography achievement levels in order to determine whether they would exhibit better internal consistency or a better match to external criteria than the 1992 reading or mathematics achievement levels.[6] In fact, the Panel once again found troubling differences in achievement-level cutscores set using dichotomous versus partial-credit (extended-response) items. *Although not as dramatic as the differences found for the 1992 achievement levels, the results again showed that levels set using extended-response items were considerably higher than those set using multiple-choice or dichotomously-scored constructed-response items.* These differences were evident despite the fact that the process of translating item judgments into cutscores automatically takes account of differences in item difficulty.

The Panel also noted that only 12 percent of 12th graders were classified as at or above the proficient level for U.S. history, compared to 29 percent so-classified for world geography. These results, which are counterintuitive in view of the fact that virtually every high school student in the country takes U.S. history whereas very few take world geography, suggest that there is not a consistent frame of reference for achievement expectations when levels are set in different subjects. The discrepancy between the history and geography results was discussed in detail by the Governing Board when it was deciding whether or not to release the 12th-grade U.S. history results; however, no mention of the anomaly was made when the results were presented to the public.

Finally, in an attempt to determine whether the achievement-level cutpoints continue to be set too high, the Panel examined performance on the AP examination in U.S.

[6] U.S. history and world geography were assessed nationally in 1994, but were not included in the TSA. It was not therefore in the Panel's purview to conduct a formal evaluation of the achievement levels set for these subjects. However, to the extent that the data were readily available, the Panel believed it should determine whether or not the results from these new level-setting efforts confirmed the Panel's earlier findings.

history.[7] Many colleges and universities give college credit for AP courses taken in high school if students score three or better, and the Panel found that 2.8 percent of the country's high school seniors met this criterion on the AP U.S. history examination. By contrast, NAEP classified only 1 percent of high school seniors at the advanced level in this subject. Moreover, the percentage passing the AP criterion would be even higher if AP programs were available in all U.S. high schools instead of only half of them. These findings provide additional evidence that the Governing Board's achievement levels are set too high, that is, that the achievement levels identify fewer 12th graders as advanced than actually are performing at an advanced level.

Based on its accumulated evidence, the Panel believes that NAGB should 1) institute a competition for the design of new methods for setting performance standards for all NAEP subjects with the goal of having a new methodology in place by the time of the year 2000 NAEP assessment; 2) rigorously evaluate any new standard-setting methodology *before* making it an operational part of NAEP reporting; and 3) caution readers of NAEP reports that the current achievement levels are based on a methodology that earlier evaluation panels have questioned in terms of their accuracy and validity. As a result, achievement levels results should be interpreted as suggestive rather than definitive.

## *Reporting and Dissemination*

Except for the *First Look* reading report, which was released in April, 1995, none of the reports or other data documents related to the 1994 TSA were released until March, 1996. The Panel therefore used what was learned from previous TSA reporting cycles to structure an evaluation of reporting and dissemination that was based largely on dissemination plans, reviews of prereleased copies of the reports, and consideration of the factors impinging upon the report development process.

The Panel identified four criteria fundamental to successful reporting:

- The accuracy of the results;
- The likelihood that the results will be interpreted correctly by the intended audience;
- The timeliness with which the results are made available; and
- The extent to which the results are accessible and adequately disseminated.

The Panel commends NCES and NAGB for continuing to search for data displays and report formats that are more comprehensible to the lay reader and more likely to yield correct interpretations. The 1994 reports generally showed improvements in this respect. NAEP has not done as well, however, in its efforts to speed up reporting, an outcome with high priority for much of state NAEP's constituency. In fact, the new *First Look* report, which contained only summary findings for the 1994 reading

[7] Only U.S. history could be considered because no AP examination is offered in world geography.

assessment, was not released until April, 1995 (13 months after the administration), and, although no one would have expected it at the time, the main reading reports did not appear for nearly another year after that. *This two-year lag between the assessment and the release of the 1994 Report Card was the longest in the six-year history of TSA reporting.* Factors which contributed to the delay included unexpected data problems, shifting program priorities, and competition for the services of qualified analysis staff.

The Panel shares NCES' and NAGB's concerns about the long lag time to reporting. A two-year lag is confusing to the public and clearly unacceptable when the program is back out in the field collecting the next round of data before the results are released. Nevertheless, care for the accuracy of the data must take precedence over other considerations and, given the analysis problems that occurred in 1994, the best and only choice was to delay the reports until matters had been investigated and, where possible, corrected.

The Panel concluded its discussion of reporting issues with several suggestions for improving accessibility and dissemination. These included 1) enlisting media representatives to help identify the most comprehensible methods for displaying the data; 2) preparing additional focused research reports for various audiences, including reports that draw on other research, where possible, to corroborate and inform the correlational relationships observed in the NAEP data; 3) providing more examples of assessment tasks and student responses; and 4) exploring ways to support states in generating their own reports of NAEP findings.

## *Utility of the TSA*

The final perspective that bears on the overall evaluation of the TSA, and in effect subsumes all other perspectives, concerns its utility. As suggested above, utility must rest, firstly, on the validity and reliability of the data. Beyond this, the results must be timely, accessible, and policy relevant, and the program must be perceived as useful and valuable by the major customers of the information it provides—particularly the states. To investigate the latter, the Panel commissioned surveys and case studies of NAEP's perceived influence after the release of each round of TSA data, concluding with a set of case studies and a mail survey of state assessment directors, mathematics specialists, and reading specialists in December, 1995. Throughout its evaluation, the Panel also monitored media coverage of NAEP and the TSAs and followed the opinions and actions of other NAEP stakeholders.

### *Utility of NAEP Data to the States*

For the most part, the Panel concluded from these efforts that state NAEP has become a valued indicator of educational progress and has served particularly to provide an independent validity check on the states' own assessments. The latter role has been especially important during a period when many state assessments have undergone radical reform, making upward or downward trends in results particularly difficult to

interpret. In Rhode Island, for example, the state reading specialist reported that the 1994 TSA reading results provided important evidence for the success of an ongoing reading initiative. The fourth-grade class that took the 1994 TSA was the first cohort to go through grades K-three after implementation of the state's Literacy Act, which addresses language arts instruction. Their performance was better than that of the cohort that took the TSA in 1992, although the change in average TSA proficiency scores was not statistically significant between the two years. The fact that the TSA results confirmed a similar upward trend in scores on the state's own performance-based writing assessment, however, made the evidence from both sources more convincing.

Thus, whereas state NAEP cannot and should not replace individual state assessments that are better tailored to state goals and curricula and offer more comprehensive coverage of schools and students, several factors contribute to NAEP's value as an external monitor. The assessment's forward-looking content and format, its secure status,[8] and the rigorous statistical standards maintained in data collection, analysis, and reporting all contribute to the credibility of cross-state and cross-year comparisons generated by the TSA.

When state NAEP results have yielded dramatic or unexpected results, particularly when a state's students performed worse than expected, considerable public debate has followed. North Carolina and California both provide notable, and very different, examples of this effect. In 1990, North Carolina educators were dismayed to discover that the state's students had done much worse on the NAEP mathematics assessment than the educators had expected, based on results from the state's own assessment, a commercially available, norm referenced test. During the subsequent debate and discussion, decision makers concluded that their current state assessment did not cover a full range of valued mathematics outcomes and that classroom teachers typically lacked certain key understandings required to successfully implement the forward-looking mathematics curriculum that North Carolina had recently introduced. Intensive in-service training, based in part on materials and data from the 1990 NAEP, was undertaken, and North Carolina experienced a significant gain in eighth-grade mathematics achievement between the 1990 and 1992 TSAs.

In California, educators and the public were also shocked when fourth-grade reading achievement estimates from the 1994 TSA showed California performing significantly worse than it had done in 1992, and positioned virtually at the bottom of the distribution of participating states.[9] This information was particularly important in view of the fact that California's own assessment system has been in disarray for the past several years, precluding any meaningful assessment of performance trends from that particular source. In the resultant furor, most commentators simply claimed the TSA results as further evidence for what they already felt was wrong with the state's education system, whether that was crowded classrooms or the state's whole language

[8] Because NAEP test booklets are kept secure between assessments, the program minimizes inappropriate "teaching to the test" and the upward score drift so often observed in results from states or commercial assessments.

[9] In absolute numbers, only Louisiana had lower mean 1994 reading results than California. When one considers the statistical significance of the differences however, California's performance was approximately equal to that of seven other states, including Louisiana, at the bottom of the distribution.

reading curriculum. (In fact, the California curriculum, which may or may not be well implemented, is a variant on a well established reading model that has proven successful in other states and is compatible with the framework underlying the NAEP reading assessment.)

About 60 percent of the states that undertook revisions to their mathematics or reading curricula during the past five years reported NAEP as a notable source of ideas. Similar numbers referred to NAEP as a model, or a source of external validation, for changes to their reading or mathematics assessments. State educators, for example, have closely followed NAEP's pioneering efforts to set performance standards, and both assessment directors and curriculum support staff have used NAEP's external credibility to argue for such desired objectives as better alignment with National Council of Teachers of Mathematics (NCTM) standards, and reading models based on reading for meaning, higher order skills, and real-world reading tasks. Furthermore, some states have been especially proactive in tailoring NAEP's published materials to use with local educators. After the 1990 assessment, for example, the North Carolina state department of education made extensive use of NAEP frameworks and example items to provide in-service training for district and school staff. As noted above, these efforts were largely motivated by North Carolina's poor showing on the 1990 TSA, and the in-service materials were made particularly salient by incorporating specific information about how children in North Carolina had performed on the example items and about how North Carolina's teaching practices and teacher preparation compared with national averages.

## *Contributions to the National Debate*

Interestingly, state NAEP has broadened NAEP's influence not only at the state level, as might be expected, but also at the national level. NAEP has been adopted by the National Goals Panel as the primary indicator of progress towards goal three, which states that

> By the year 2000, American students will leave grades four, eight, and twelve having demonstrated competency in challenging subject matter including English, mathematics, science, history, and geography; and every school in America will ensure that all students learn to use their minds well, so they may be prepared for responsible citizenship, further learning, and productive employment in our modern economy.[10]

NAEP also routinely receives national press coverage after each of its major data releases. The latter has tended to be more widespread when regional media are able to tie in results for their own states. Publications devoted to education news, such as *Education Week,* also contain frequent references to NAEP, both as a unique source of information about education achievement and as a model for current assessment practices. Finally, for better or for worse, NAEP has recently found itself spotlighted by requests from two other federal programs (the National Science Foundation's

[10] National Education Goals Panel, *The National Education Goals Report: Building a Nation of Learners* (Washington, D.C.: Author, 1991), 10.

Systemic Initiatives in mathematics and science, and the Title I program for disadvantaged students) that were considering the use of NAEP instruments in their evaluations.

### *Limitations on State NAEP Utility*

Despite these many positive findings, many states also reported factors that limited the utility of NAEP results or threatened school participation rates. The most common concerns noted in the Panel's surveys and case studies were the long lag time to reporting and the fact that NAEP provides no district- or school-level results. The staff time required to coordinate the assessment, the unpredictability of the assessment schedule, and the burden on participating schools—particularly among smaller or less densely populated states—were also mentioned with some frequency.

Much remains inconclusive with regard to how the costs and benefits of state NAEP will play out in states' willingness to participate over time. For example, some states made early decisions to use NAEP as a component of their own state testing programs or to fill special needs—for example, the need for baseline or follow-up measures of achievement to use in evaluating the impact of programmatic interventions. In these instances, the unpredictability of the NAEP assessment schedule has proven particularly problematic because states found that they had allocated portions of their assessment budgets without getting back the information they had anticipated.[11] If states could be certain about exactly what state NAEP would offer several years in advance, they might be willing to transfer more of their state assessment budgets to this purpose, and perceptions of burden might decrease. Alternatively, if state NAEP begins to assess two subjects and two grades per year, as NAGB is currently considering, a number of states might find participation to be too burdensome under any circumstances. Predicting either of these outcomes can only be speculative however, and factors external to NAEP may weight the consequences more heavily than anything NAEP itself could do. For example, if the current emphasis on standards and accountability through assessment is sustained, NAEP may be relatively more valued, whereas if the recent emphasis on state and local control grows stronger (with or without an emphasis on standards), more states could draw back from a federal testing program.

## *The Impact of State on National NAEP*

When state NAEP was authorized by Congress on a trial basis in 1988, one of Congress' central concerns was whether state NAEP would have a deleterious effect on national NAEP. By asking this question, Congress was tacitly affirming the importance of protecting the integrity of national NAEP and expressing a concern that state NAEP might have a negative impact on state participation in national NAEP, especially in the case of small states.

[11] For instance, states may have invested in 1990 NAEP to get a pre-measure in mathematics, only to find that they would not be able to get a post-measure later (e.g., in 1994) when they needed it.

In examining the answer to this question in its evaluation of the 1990 TSA, the Panel answered Congress as follows:

> Although it is true that state NAEP increased the burden on small states to provide sufficient numbers of schools to meet the requirements of both the national and state samples, few of the 37 states that participated in the first trial found this to be a problem. Moreover the 1990 national NAEP school cooperation rates were similar to the cooperation rates obtained in the other recent NAEP administrations, and the 1990 TSA school cooperation rate was higher. It is of significance as well that the administration of the 1990 trial did not interfere in any way with the administration of national NAEP. There is, therefore, no indication at this time that a program of state NAEP testing will have a negative impact on the national indicator system. However, the Panel recognizes that this will have to be reevaluated after the 1992 trial in which the testing burden will have increased due to the addition of assessments at grade four in reading and mathematics.[12]

The Panel noted a considerable drop in the initial school participation rates (percent of originally sampled schools that actually participated in the assessment ) as part of its evaluation of the 1992 TSAs. The 1990 initial school participation rate was 94 percent. By comparison, the participation rate for the 1992 trials in both the fourth- and eighth-grade trials was 88 percent. Although the Panel expressed some concern about the drop, it acknowledged that the 88 percent rate for the 1992 TSAs still compared favorably with the initial school participation rates for national NAEP, both in 1992 and earlier. Furthermore, the 1990 rate of 94 percent might be explained as at least partially due to the "novelty" of being able to participate in the very first trial state NAEP.

The Panel did express concern, however, that the number of states with before-substitution participation rates below 85 percent had dramatically increased between 1990 and 1992. In 1990, only two states failed to meet this criterion; in 1992, 17 states fell below 85 percent on one or more of the trials. The Panel attributed at least some of this increase to the burden the states had to bear in moving from one trial in 1990 to three trials in 1992. Though the Panel recommended that state NAEP be continued, it also expressed concern about the burden that a three-grade trial in multiple subjects (as was being considered at the time) would create for the states. Partly for this reason, the Panel recommended that the evaluation of state NAEP continue.[13]

The number of states participating in the 1994 trial was the same as in the 1992 trial (44 states and territories). The initial school participation rate for the 1994 TSA was actually one percent higher (89 percent), and the number of states with before-substitution participation rates below 85 percent dropped to 12. Whereas this was good news, it must be remembered that there were three state NAEP trials in 1992, (eighth-grade mathematics, fourth-grade mathematics, and fourth-grade reading) and

[12] The National Academy of Education, *Assessing Student Achievement in the States* (Stanford, CA: Author, 1992).

[13] Ibid., 104.

only one in 1994. Therefore, at least some of the improvement in participation must be due to the decreased burden of the smaller assessment.

Most importantly, there is only a small amount of evidence that the burden of participation in state NAEP is having any impact on participation in national NAEP. The move from a single trial in 1990 to three trials in 1992 was accompanied by a decline in the national NAEP initial school participation rates from 87 percent to 85 percent for grade eight and 88 percent to 86 percent for grade four. For 1994, the initial participation rates for both grades in the national NAEP remained approximately constant, at 86 percent. What the Panel cannot know, however, is what the 1994 national participation rates would have been had there been state NAEP trials of two subjects at two grade levels, as had been planned until budget limitations forced the reduction to a single trial at one grade.

The Panel believes that an implicit, mostly unspoken *quid pro quo* has developed between the states and NAGB, by means of which the states are willing to participate in national NAEP *at least in part* because of the value they get from participation in state NAEP. Since 1990, the Panel has observed movement from guarded cooperation among participating states to general anticipation when state NAEP results are about to be released. Positive attitudes toward state NAEP can only grow if Congress and the Governing Board are able to guarantee a firm state assessment schedule that ties down the subjects and grades to be assessed several years in advance. As a result, the Panel suspects that if Congress were to recommend the abandonment of state NAEP at any time in the near future (an action which the Panel believes almost certainly will not happen), the motivation for states to continue in national NAEP could drop precipitously. The discussions at Governing Board meetings also suggest that NAGB regards state NAEP to be at least as, if not more, important than national NAEP, despite the fact that state NAEP is still, by law, developmental, and national NAEP is not.

*As a result, in contrast to its original conclusion at the end of the evaluation of the 1990 TSA, which was simply that state NAEP had had no deleterious effect on national NAEP, the Panel now believes that the fortunes of the two programs have become increasingly intertwined.* State NAEP has greatly increased the visibility and perceived utility of the entire NAEP program, and suggestions for merging the state and national samples continue to arise (although it is not evident that such a merger would be feasible or significantly reduce burden). Monies spent on state NAEP obviously also interact with monies available to maintain a quality national NAEP program, although the nature of this interaction is complex. On the one hand, the substantial funds spent for state NAEP cannot then be spent for other NAEP activities. On the other hand, the heightened visibility conferred by state NAEP may result in a net increase in national NAEP resources. For example, the substantial framework and item development efforts that have characterized the last several years have benefited both programs and might not have been funded without the impetus of state NAEP.

## *The Panel's Recommendation for the Continuation of State NAEP*

Often it takes fewer words to affirm that something was done well than to point out those issues and concerns that nevertheless remain. The various points identified

throughout this report for further review and study should not detract from the overall success of the Trial State Assessments. Based on its evaluation of the TSAs, the Panel concludes that state NAEP has been shown to be a valid, reliable, and useful measure of student achievement, and that it aligns favorably with the Panel's *quality, utility, and state indicator principles*. For these reasons, the Panel recommends that state NAEP be continued, and that it be moved from developmental to permanent status when NAEP is next reauthorized. However, in light of its size and cost, the Panel further recommends that the scope and function of state NAEP be reviewed regularly, and particularly after any substantial change in mission or design. Such re-evaluation should be done in the context of the overall NAEP program and with the abiding aim of providing the best and most useful information about student achievement for the nation.

Some of the issues that should continue to be examined in the future include

- The viability of continuing to assess nonpublic schools in the state NAEP program;
- The value and feasibility of grade 12 state assessments;
- The tension between including as many students with disabilities and limited English reading skills in the assessment as possible, and the cost of doing so;
- The adequacy of the present NAEP design to meet the increasing demands of NAEP's stakeholders while still satisfying the Panel's *quality principle;* and
- The continuing use of achievement levels to report NAEP results.

With regard to the latter, the Panel understands and endorses the value of reporting results against rigorous standards. Because of the problems of reliability and validity affecting the current achievement levels however, research and development for new performance standards should begin as soon as possible.

The panel looks forward to presenting its capstone report in the summer of 1996. The capstone report will review the findings, recommendations, and continuing issues from the Panel's evaluations of the 1990, 1992, and 1994 Trial State Assessments. In addition, the report will look forward to the year 2000 and beyond, considering recommendations for the design of a NAEP program that offers quality assessments for the nation and the states and also anticipates the changing nature of educational practice as the latter will be influenced by technology and by our developing knowledge of learning and human cognition.

# Appendix A

## *Detailed Scoring Guides and Examples of Student Responses for Sample Assessment Items Shown in Figure 2.1*

### SCORING GUIDE

*Item 3. Do you think Turtle should have done what he did to Spider? Explain why or why not.*

**4 = Acceptable**

Acceptable responses either agree or disagree with Turtle's actions, and mention something about the fact that Spider was mean to Turtle or he tricked him and Turtle got revenge.

The information included in the explanation must be appropriate within the context of the story. For example:

- "Turtle was right to get back at Spider because Spider was mean to him."
- "Turtle was not right to get back at Spider because it is not a nice thing to do."

**1 = Unacceptable**

- "Turtle was right to get back at Spider because Turtle was hungry."

### ADDITIONAL NOTES FROM SCORER TRAINING

A 4-level response gives an opinion and...

**YES**

- implies Spider did something bad, wrong (Spider was guilty)
- suggests Spider deserved it
- implies Turtle was getting revenge (getting back at, doing same thing)
- suggests that Spider needed to learn a lesson

**NO**

- suggests treating others the way you want to be treated
- implies that it is against the student's morals/upbringing

Examples of 4's

- Yes, Spider was bad/mean/selfish
- Yes, because Spider did wrong
- I think Spider deserved getting tricked because he was so selfish (implies "yes")
- No, then he'd be as bad as Spider

Examples of 1's

- Yes/no because Spider/Turtle was wrong (needs explanation)
- Yes/no, Turtle was nice (needs explanation)
- No, Spider was bad
- It was mean
- He was mean
- No, it would be a mean thing to do
- No, because that's very rude
- I don't think so, because he's mean
- No, because it was out of character for Turtle (this is unacceptable because Turtle did do it; it was his character.)
- No, because he ate all of the food, and he didn't leave Spider any

NOTE: The full guide included four to six examples of student responses at each level. Two of each are reproduced here.

## EXAMPLES OF STUDENT RESPONSES TO ITEM 3

**Level 1 (Unacceptable)**

10. Do you think Turtle should have done what he did to Spider? Explain why or why not. WO000048

He shoud do that he is mean.

10. Do you think Turtle should have done what he did to Spider? Explain why or why not. WO000048

No becoue turtle is very hungry.

**Level 4 (Acceptable)**

10. Do you think Turtle should have done what he did to Spider? Explain why or why not. WO000048

I think he sould not of because to wrong's don't make aright.

10. Do you think Turtle should have done what he did to Spider? Explain why or why not. WO000048

Yes, because he wanted to get evan.

*Item 4. Think about Spider and Turtle in the story. Pick someone you know, have read about, or have seen in the movies or on television and explain why that person is like either Spider or Turtle.*

**Stance**
Personal Response

**General Scoring Rubric**
Demonstrates comprehension of the character of Spider or Turtle and relates the qualities of a story character to a person or character about which the respondent is familiar.

**Scoring Rationale**
This task requires students to:

- provide evidence that they understand the character of Turtle or Spider:
- relate the properties of Spider or Turtle to a similar person or character; and
- explain the similarities between Turtle or Spider and a similar person or character.

**1 = Unsatisfactory** These responses demonstrate little or no understanding of the character of Spider or Turtle. In these responses, students often name a character, but do not relate this character to Spider or Turtle, or provide only inappropriate characteristics. Also, students may summarize a story or movie but do not relate it to Spider or Turtle in any obvious way.

For example:

a. "My friend is like Turtle."

b. "Turtle always washed his feet."

c. "My friend Jeff because he is friendly."

d. "Someone in my class is friendly like Spider."

**2 = Partial**

These responses indicate some understanding of the character of Spider or Turtle in the story by providing information about the character of Spider or Turtle, but fail to make a connection with a real world person or character; or they make the connection between a story character and a real world person/character, but don't distinguish if the story character is like Spider or like Turtle.

For example:

a. "Turtle was able to get revenge on Spider for the way that Spider treated him."

b. "My friend is like Turtle because they both wipe their feet a lot."

c. "Mindy is like Turtle because she is friendly "

d. "On Chip and Dale's Rescue Rangers. Dale is always selfish."

**3 = Essential**

These responses demonstrate a good understanding of the character of Spider or Turtle by providing any important character trait that is related or linked to a real world person or character.

For example:

a. "My older brother and Turtle are alike because they both get revenge on their enemies."

b. "My sister is like Spider because she likes to trick people."

**4 = Extensive**

These responses demonstrate an in-depth, rich understanding of the character of Spider or Turtle and link this understanding to a real world person/character. Evidence of depth of understanding includes describing more than one essential story character trait linked to a real world person/character, providing a sophisticated interpretation of an essential story character trait that is linked to a real world person/character, identifying how a real world character is like Spider in one way and like Turtle in another, or identifying a pair of real world characters and explaining how these two characters are like Spider and Turtle.

For example:

a. "My brother and Spider are alike because they both unfairly control the people around them to their own advantage."

b. "My brother and Spider are alike because they both cheat people and are selfish."

c. Scrooge is like Spider because he is greedy and Bob Cratchet is like Turtle because he gets something in the end too.

d. "My friend Anne is like Turtle because when someone calls her names she just walks away, but sometimes she's like Spider and plays tricks on people."

## Additional Notes From Training

Equations for four-level responses:

| | | | | |
|---|---|---|---|---|
| A person | + | Spider | + | 2 important characteristics of Spider |
| A person | + | Turtle | + | 2 important characteristics of Turtle |
| A person and Another person | + | Spider and Turtle | + | 1 important characteristic of Spider and 1 important characteristic of Turtle |
| A person | + | Spider and Turtle | + | 1 important characteristic of Spider and 1 important characteristic of Turtle |

## TURTLE

| WRONG | UNIMPORTANT | IMPORTANT |
|---|---|---|
| generous | trusting | doesn't speak up |
| sharing | nice | quiet |
| honest | friendly | doesn't complain |
| dependable | kind/caring | obedient |
| gets mad | tired | will not tell on you |
| helpful | slow | polite/proper/good manners |
| not greedy | good | self-controlled |
| dumb | has dirty feet | respectful |
| suspicious | lives in water | considerate |
| mean | never hurts anyone | hardly ever hurts anyone |
| was not welcome | afraid to speak up | patient |
| learned a lesson | sweet/loving | doesn't get mad |
| understanding | stingy/greedy | doesn't fight |
| can talk | selfish | easy going |
| | not afraid of spider | puts up with things |
| | doesn't like to miss meals | smart/sly |
| | eager | clever/tricky |
| | not mean | outsmarts spider |
| | good | gets back/even |
| | shy | teaches a lesson |
| | organized | gets revenge |
| | invites someone for dinner | tries over and over |
| | do the same to each other | persistent |
| | doesn't get fed | easily tricked at first |
| | clean | suckered |
| | breaks promise | trusted spider |
| | fair | hungry/desperate for food |
| | | ashamed |
| | | doesn't want people to know he was tricked |

## SPIDER

| WRONG | UNIMPORTANT | IMPORTANT |
|---|---|---|
| steals | trusts | two-faced |
| doesn't trust | picky | gets tricked |
| learned a lesson | orderly | recipient of revenge |
| friendly | eats a lot/a pig | greedy |
| eager | always hungry | selfish/wants own way |
| sharing | eight legs | always gets what he wants to eat |
| bragging | very light | doesn't share |
| kind | wears a jacket | always wants things for himself |
| teases | makes people do things | let's turtle starve |
| makes fun of | people avoid him | doesn't have feelings for others |
| plays jokes | bad attitude | not polite |
| tries to steal a scene | invites someone for dinner | mean |
| gets mad | bossy | not nice |
| doesn't like people | bully | not kind |
| lonely | picks on people | not friendly |
| sets traps | hurts people | bad manners |
| likes to do things people don't like | self-controlled | hurts people's feelings |
| can talk | doesn't get food at someone's house | treats people badly |
| smart aleck | impatient | bad/unfair |
| lies | gets what he deserves | rude/evil |
| | sleazy | cruel |
| | breaks promises | tricky |
| | makes people do things over and over | cheats/sly/smart |
| | makes people do things they don't want to do | clever |
| | is talked about behind his back | |
| | is concerned about what people think of him | |

NOTE: The full guide included four to six examples of student responses at each score level. Two of each are reproduced here.

## EXAMPLES OF STUDENT RESPONSES TO ITEM 4

**Level 1 (Unsatisfactory)**

6. Think about Spider and Turtle in the story. Pick someone you know, have read about, or have seen in the movies or on television and explain how that person is like either Spider or Turtle. WO000043

My friend Sacha likes turtles she likes when these head pop in and out there is also big turtle and small turtles Snapping turtle too. Some turtles are as big as 3 feet wide and 4 feet long some even has dots on them and my mom when she was about nine years old she went down to a pond and caut them then let them go she named one Lisa too.

6. Think about Spider and Turtle in the story. Pick someone you know have read about, or have seen in the movies or on television and explain how that person is like either Spider or Turtle. WO000043

Daffy Duck, becase he ~~does~~ does always tries to steal a sea

**Level 2 (Partial)**

6. Think about Spider and Turtle in the story. Pick someone you know, have read about, or have seen in the movies or on television and explain how that person is like either Spider or Turtle. WO000043

Peggy on Married With children She is selfish and lets her children starve.

Level 2 (cont.)

6. Think about Spider and Turtle in the story. Pick someone you know, have read about, or have seen in the movies or on television and explain how that person is like either Spider or Turtle. WO000043

Well once I saw a movie and this person was so selfish that he wouldn't anyone touch his stuff ever and everyone called him a selfish person.

Level 3 (Essential)

6. Think about Spider and Turtle in the story. Pick someone you know, have read about, or have seen in the movies or on television and explain how that person is like either Spider or Turtle. WO000043

My brother is greedy like spider.

6. Think about Spider and Turtle in the story. Pick someone you know, have read about, or have seen in the movies or on television and explain how that person is like either Spider or Turtle. WO000043

In Tom and Jerry cartoons, Tom can be like Spider. Tom can seem to be acting friendly. Then it can turn out that it was a trick. If Tom puts cheese down, it is usally on a mouse trap.

6. Think about Spider and Turtle in the story. Pick someone you know, have read about, or have seen in the movies or on television and explain how that person is like either Spider or Turtle. WO000043

Scrug is a lot like the spider because thay are both greedy but the spider wants food and Scroge likes money. and Bob cratchet was kind of like the turtle cause he gets somethin in the end to

6. Think about Spider and Turtle in the story. Pick someone you know, have read about, or have seen in the movies or on television and explain how that person is like either Spider or Turtle. WO000043

MY Friend Anne is like turtle. If someone calls her Names she just walks away. But sometimes shes like spider and plays tricks on people. But shes Mostly like turtle she honest and dependable. she even gets along with brothers. (Most of the time)

# *Appendix B*

## *Reading Experts Participating in the Panel's Content Validity Study for the 1994 TSA*

**PRINCIPAL INVESTIGATORS**

David Pearson
*Michigan State University*

Lizanne DeStefano
*University of Illinois*

Peter Afflerbach
*University of Maryland*

**PANELISTS**

Diane Bottomley
*Pennsylvania State University*

Marsha DeLain
*Delaware Department of Public Instruction*

Jan Dole
*University of Utah*

Eunice Greer
*Harvard PACE*

Greg Morris
*University of Pittsburgh*

Susan Neuman
*Temple University*

Charles Peters
*Oakland Schools, Waterford Michigan*

Suzanne Wade
*University of Utah*

Arlette Willis
*University of Illinois*

# Appendix C

*Synopses of Studies for The National Academy of Education Panel on the Evaluation of the National Assessment of Educational Progress Trial State Assessment*

This appendix contains summaries of 10 studies that were commissioned by the Panel and served as the basis for the Panel's findings and recommendations. The complete reports are being published in a separate volume, *Quality and Utility: The 1994 Trial State Assessment in Reading, Background Studies*.

# *Content Validation of the 1994 National Assessment of Educational Progress in Reading: Assessing the Relationship Between the 1994 Assessment and the Reading Framework*

Lizanne DeStefano
*University of Illinois at Urbana-Champaign*

P. David Pearson
*Michigan State University*

Peter Afflerbach
*University of Maryland at College Park*

## *Introduction*

This paper presents the fourth in a series of studies commissioned by the Panel to examine the content validity of the NAEP reading assessment. Previous studies evaluated the validity of the framework and the match between the framework and the 1992 assessment items. The present study considered the quality of the items used in the 1994 reading assessment and the extent to which the items, as a whole, matched the scope and substance specified by the reading framework.

The framework was developed in 1991 and will remain in place until the next revision authorized by NAGB. Written over an 18-month period by a steering and a planning committee composed largely of professional educators, the framework serves as the basis for all national and state NAEP reading assessments. It specifies an assessment based on 1) a constructivist view of comprehension and learning; 2) the use of longer reading passages based on authentic texts; 3) an increased emphasis on constructed-response items; and 4) a reading domain defined by a two-dimensional matrix of situation by stance. Situation takes account of purpose for reading and type of text, while stance refers to the manner in which the reader interacts with the text.

The three reading situations, which form the basis for reporting subscales as well as defining the overall scope of the assessment, are 1) reading for literary experience; 2) reading for information; and 3) reading to perform a task. At grade four, only the first two situations are included in the assessment. The four reading stances, each of which applies to every reading situation, are 1) forming an initial understanding; 2) developing an interpretation; 3) personal reflection and response; and 4) demonstrating a critical stance.

## *Method*

Nine expert panelists (three at each grade level assessed by NAEP) were convened for a two-day session to review the items, reading passages, and scoring criteria used in 1994. The panelists classified the items according to the matrix defined in the framework, then addressed three questions related to the content validity of the assessment: 1) how well did their classifications replicate the classifications assigned

by the NAEP item developers? 2) how well did the distribution of items match the distributions specified by the framework? and 3) how appropriate were the scoring guides for the constructed-response items? In addition, the panelists were asked to develop suggestions for future validity studies of NAEP.

Although the official NAEP classifications only identify a single situation and a single stance for each item, panelists were allowed to designate both a primary and a secondary stance. This strategy was included to alleviate the difficulty that participants in the 1992 validity study reported when they were required to assign each item to a single stance. The underlying difficulty reflected an apparent shortcoming of the framework due to the fact that it assumed mutual exclusivity across stances in cases in which, in fact, items contained characteristics of more than one stance. The panelists' primary classifications, however, are used in the findings reported here.

## *Findings*

### *Agreement on NAEP Classifications*

After the three panelists at each grade level worked together to reach consensus about the classification of the items, these consensus classifications were compared to the "official" NAEP classifications. Averaged across all grade levels, agreement was 98 percent for reading situation. For reading stance, agreement rates ranged from 61 percent for 4th grade to 70 percent for 12th grade for primary classifications. Across all three grade levels, agreement on stance was highest for developing interpretation (79 percent), followed by initial understanding (71 percent) and personal reflection and response (71 percent). Items classified by NAEP as demonstrating a critical stance exhibited the lowest levels of agreement for all grade levels (44 percent).

When the panelists were asked to explain their disagreements, they often referred to ambiguities in the definitions of the reading stances and the lack of clear examples presented in the framework. They also identified between-stance ambiguity and overlap throughout the framework. Another source of disagreement stemmed from discrepancies between the intent of an item and the way in which it was scored, as expressed by the scoring guide. For example, panelists were critical of scoring procedures for constructed-response items that appeared to assess critical stance, yet required students to give examples from the text.

The authors identified a pattern of mismatches in the disagreements between the panelists' and the official NAEP classifications that they conceptualized as follows. First, the authors proposed the hypothesis that the relationship between the reader and the text becomes more abstract and reflective as the reader progresses from initial understanding, through developing interpretation, then personal response, and finally, to critical stance. For example, whereas critical stance requires judgments and objectivity, initial understanding asks only for an initial impression of what was read.

Within this context, the authors identified a consistent tendency for the panelists to view the items as "closer to the text" and less reflective than did the NAEP item developers. For example, many items that were officially classified as critical stance were categorized by the panelists as developing interpretation. DeStefano et al.

concluded that the panelists did not see the same set of logical, cognitive, and rhetorical relations in the items as did the item developers. In the cases in which differences in judgment occurred, the panelists took a more constrained view of the challenge involved.

### *Determining the Fit of the 1994 NAEP Assessment with the Reading Framework*

The panelists next examined the extent to which the overall item pool for the 1994 NAEP assessment matched the specifications laid out by the reading framework. With regard to reading stance, the specifications call for testing time at each grade to be divided into thirds, with one-third devoted to initial understanding and developing interpretation, one-third spent on personal reflection and response, and one-third assigned to critical stance. As a consequence of their systematic disagreements with the official NAEP classifications by stance however, panelists found that the item pool for the 1994 assessment devoted markedly less time to items addressing personal reflection or critical stance. This deficiency was most pronounced at the fourth grade, in regard to which panelists felt that only 8 percent of assessment time was spent on items that assessed critical stance, whereas the overwhelming majority of fourth-grade assessment time was spent on understanding and interpretation.

With regard to reading situations, the intended distributions put forth in the framework were as follows:

| | Percent of Items | | |
|---|---|---|---|
| | Reading for Literary Experience | Reading for Information | Reading to Perform a Task |
| Grade 4 | 55 | 45 | 0 |
| Grade 8 | 40 | 40 | 20 |
| Grade 12 | 35 | 45 | 20 |

Although the panelists agreed with the "official" NAEP item classifications for reading situation, they found that the assessment item pool gave somewhat greater emphasis to reading to perform a task in the 8th (27 percent) and 12th (27 percent) grades than called for in the framework. They did not, however, consider this level of discrepancy to be a significant problem. Of greater concern was the validity of the distributions *specified* in the framework. Specifically, panelists questioned whether the distribution across reading purposes called for in the framework actually reflected the types of reading that students typically do in schools, as was intended. Panelists were particularly critical of the omission of reading to perform a task at the fourth grade, feeling that this incorrectly implied that fourth graders do not read for this purpose. Further, although they generally agreed that the framework distributions adequately reflected reading purposes at the middle and secondary levels, panelists felt that there

could have been sharper increases in the numbers of items addressing reading for information and reading to perform a task as grade levels increased.

***Critique of the Scoring Guides for Constructed-Response Items***

Panelists identified four problematic aspects of the scoring guides:

> First, an implicit "more is better" strategy in which, for example, scores were determined by the *number* of reasons proffered by the student, rather than other indicators of reading skill such as depth of understanding, complexity of thought, or recognition of relationships.
>
> Second, in a number of instances, a discrepancy between the intent of the item (as represented by the NAEP item type classification) and the way in which a particular scoring guide was operationalized. For example, panelists identified items that had been classified as critical stance but, based upon the scoring guides, really asked the student to develop interpretations.
>
> Third, ambiguities and inconsistencies arose in cases in which, for example, items appeared to ask for a particular reading stance, but scoring guides gave equal credit for other stances. The general form of the scoring guides that led to this problem was as follows: "A response of three can be earned either by listing two facts explicitly stated in the text *or* by providing a single new insight that is consistent with but not stated in the text."
>
> Finally, the panelists were concerned about a lack of clear communication to students about what is expected from their responses, whether that is listing three or four facts, or giving new insights about the text. Panelists noted that there is no provision in NAEP (or most assessments) for educating students in advance about the criteria that will be used to score constructed responses.

***Suggestions for New Validity Initiatives***

The panelists suggested a number of additional validity studies to improve the content of the NAEP reading assessments. First, the panelists want a more thorough construct validation of the NAEP reading framework to determine if it captures the actual processes of reading engagement by real readers. Such a study should encompass readers of different abilities, cultural dispositions, and backgrounds who are asked to read a broad sampling of real texts. Panelists felt strongly that such a study would give both reading experts and NAEP greater confidence in the particular direction taken in the current framework.

The panelists also suggested that a careful evaluation of the instructional validity and the policy relevance of the NAEP reading framework be undertaken. These evaluations would examine the elements of the framework in relation to both typical and exemplary instructional practices in schools and the most recent state frameworks.

Finally, the panelists suggested that a set of contextualized validity studies be utilized to examine the validity of the current NAEP reading assessment for particular groups of readers defined by race, gender, language, ethnicity, income, prior achievement, or education opportunity. Panelists expressed many concerns with respect to the

assumption that the current NAEP provides a fair and equitable assessment for the full range of diverse populations in the United States. Also of concern were cultural and linguistic differences in the interpretation of questions and directions.

### *Recommendations for Future NAEP Reading Assessments*

The panelists made a number of recommendations for consideration in the design of future assessments. Specifically, they recommended that

> The NAEP framework be carefully revised to better represent the complexities and subtleties of the reading process. Changes should also give more emphasis to reader variation as represented by individual characteristics and by membership in various social, economic, and cultural groups.
>
> Continual attention be given to the correspondence between the framework and the assessment itself, including both the items and the scoring guides. Matters of stance ambiguity and category overlap should be directly addressed, either by developing more precise definitions and tighter categories, or by developing a theory of response overlap (i.e., a principled way of accounting for expected conceptual ambiguities).

## *School and Student Sampling in the 1994 Trial State Assessment: An Evaluation*

Bruce D. Spencer
*Northwestern University*

### *Introduction*

This study represents the third evaluation of school and student sampling in the NAEP TSA carried out by the author on behalf of the Panel. Evaluations of the 1990 and 1992 TSAs found that the samples were well designed and implemented. However, they also highlighted the fact that certain groups of students were excluded from the survey. In 1992, the Panel recommended that the population definition for the TSAs be broadened to include students in private schools and that the appropriateness of exclusion decisions regarding LEP and IEP students be examined in greater detail.

By 1994, a number of changes had been made to the TSA sampling procedures, and the Panel directed the author to investigate the effects of these changes as part of the overall evaluation of sampling for the 1994 TSA. This report discusses these changes, reviews the school and student sampling, weighting procedures, and participation for

the 1994 TSA, and concludes with recommendations for improved sampling procedures for the 1996 TSA.

## *Target Population*

The target population of the 1994 TSA was fourth-grade students in public and nonpublic schools in the participating states and territories. By design, those IEP or LEP students who were judged incapable of taking the test were not included, nor were home-schooled students, although charter-school students were. The validity of state achievement estimates and of state-to-state comparisons are potentially affected by between-state differences in the excluded populations or in the application of exclusion criteria.

The sampling procedures compensated for departing students (i.e., students who left the target schools between the time the sample list was compiled and the time the test was administered) by sampling a proportion of the students who moved *into* the target schools during this same time period. Unintended exclusions comprised those students absent on the day of the test and those who left the target schools and were not counterbalanced by new entrants.

### *Home-Schooled Students*

The numbers of home-schooled children are believed to be increasing, yet they still account for only approximately 1 percent of all elementary and secondary students. From a national perspective, exclusion of home-schooled children does not appear to be a significant factor in NAEP results at the fourth grade. The author therefore concluded that the effect of including home-schooled students in the TSA would be statistically small, and would be costly on a per capita basis.

### *IEP and LEP Populations*

The exclusion criteria for LEP and IEP students were part of the sampling design and were implemented locally. The local administrators were told to exclude students who could not be expected to participate meaningfully in the assessment, but to include doubtful cases. In addition, students with disabilities who were not enrolled in regular public schools were excluded. (Approximately 3 percent of all 6- to 11-year-old students with disabilities are not in regular schools.)

TSA exclusion rates were essentially the same as national NAEP exclusion rates, but the rates varied across states. Approximately 5.5 percent of students sampled for the 1994 TSA were excluded because they had individualized education plans. IEP exclusions in public schools ranged from about 2 percent in North Dakota to 9 percent in Florida and 10 percent in Maine. The IEP exclusion rates for nonpublic students ranged from about 0 percent to 1 percent for most states but did rise to 13 percent in New Mexico.

Approximately 2 percent to 3 percent of the sampled students were excluded for reasons of limited English proficiency. LEP exclusions in public schools ranged from 0 percent to 1 percent in most states, but reached 5 percent in Texas and 9 percent in California. In nonpublic schools, the LEP exclusion rates were smaller.

Although more students were excluded for IEP than LEP reasons (5.5 compared to 2.6 percent), there was more variability across states in the LEP exclusion rates (standard deviation 1.8 compared to standard deviation 1.4). Between the 1992 and 1994 TSAs in reading, the overall IEP and LEP exclusion rates for public school students remained relatively constant despite the fact that the proportions of students *identified* as LEP and IEP increased. State by state, there was strong stability in both the LEP and the combined LEP/IEP exclusion rates for grade-four reading, but somewhat less stability in the IEP exclusion rates. (The correlations between the state-level exclusion rates for 1994 and 1992 was .61 for IEP exclusions, .94 for LEP exclusions, and .84 for exclusions due to either IEP or LEP reasons.)

A key question was whether or not the exclusions were being made appropriately. Because the criterion of "meaningful participation" is vague and not operationally defined in any objective way, comparable statistics against which to validate TSA exclusion rates do not exist. One can therefore only compare TSA classification rates for IEP and LEP students against alternative figures. Nonetheless, this comparison does not address the possibility that the TSA could classify many students as IEP or LEP but exclude few, or conversely, classify few but exclude many.

## *School Sampling*

### *Public Schools*

With a few exceptions, more than 100 public schools were selected in each state. Schools that did not participate were replaced with substitute schools, resulting in a net of about 100 public schools per state. Before sampling, the state's public schools were stratified first by urbanization, and within urbanization, by minority enrollment.

### *Nonpublic Schools*

In 1992, the Panel recommended that private schools be made part of the TSA sample, the rationale being that large proportions of students are enrolled in private schools and that the states' scores and ranking could change appreciably by their inclusion. Private schools were included in the 1994 TSA, where they were combined with domestic Department of Defense schools and Bureau of Indian Affairs schools to constitute a nonpublic school sample. Approximately 15 nonpublic schools were selected from each participating state; after refusals and substitutions, net participation rates averaged approximately nine nonpublic schools per state. If fewer than six nonpublic schools in any given state agreed to participate, the nonpublic sample in the state was dropped; nine states had no nonpublic sample. Within each state, nonpublic schools were stratified for sampling first by metropolitan area status and, within metropolitan stratum, by whether the school was a Catholic school or other nonpublic school.

The 1994 TSA sampling strategy was designed to provide precise estimates for combined reports of public and private school student achievement for the states, but was never intended for separate reporting of nonpublic students. The scores were nevertheless reported separately; the only restriction on publishing was that the sample sizes be sufficient for a reliable estimate of sampling error and that participation rates for originally sampled schools meet NCES criteria. Since the 1994

administration, the NAEP contractors have recommended increasing the sample sizes for the nonpublic sector to a minimum of about 25 schools.

### *School Participation*

The participation rates for originally sampled public schools in the 1994 TSA were high for most states, with a median of 94.5 percent. The participation rates for nonpublic schools were much lower, with a median of 72 percent. For each nonparticipating school, a substitute was chosen to be as close as possible to the original school with respect to the stratification variables. In the 1994 TSA, substitution was undertaken differently than it had been in the 1990 and 1992 assessments. Specifically, the state coordinator (a state employee responsible for recruiting schools for the state sample) simultaneously received the list of original schools to be sampled and the substitute list. This practice had been avoided in previous TSAs because forehand knowledge of the substitute schools could influence the state coordinator's efforts in recruiting schools, thereby introducing bias into the assessment statistics.

### *Student Sampling and Weighting*

In each selected public and nonpublic school, a sample of about 30 students was drawn. The participation rates of the selected students for the 1994 TSA was at least 94 percent for most public and nonpublic school students. Nonpublic school students in New Mexico, with 92 percent participation, and North Dakota, with 93 percent, were two exceptions.

The majority of nonparticipating students were absentees. Each school with at least three absentees offered a make-up session, but relatively few absentee students attended these sessions. Although the weighting to reflect unequal selection probabilities was reasonably straightforward, the weighting to adjust for student nonparticipation may have been ineffective in that the only student-level characteristic that defined the weighting cells was an age dichotomy. A larger number of student-level variables such as race, sex, or Title 1 status would have been more effective for removing bias due to student-level nonparticipation.

### *Summary and Recommendations*

On the basis of his analyses, the author made a number of recommendations:

> With regard to nonpublic school participation in the TSA, nonpublic school students should continue to be included in the TSA, but their scores should not be reported separately from those of public school students, nor should scores be reported by separate types of nonpublic schools (e.g., Catholic, other religious schools, Bureau of Indian Affairs, Department of Defense).

Home-schooled students should not be included in the TSA.

More study should be given to the formulation of consistent and valid criteria for IEP and LEP exclusion and to the implementation of the criteria. If the decision to exclude some IEP and LEP students is based on the difficulty of assessing them, then the cost-benefit analysis implicit in the exclusion rules decisions should be examined.

With regard to the exclusion of IEP, LEP, and home-schooled populations, if the current inclusion/exclusion policies are continued, the reporting of NAEP results should be changed to make clear the implications of the exclusions. Even if the exclusion procedures are modified so that most or all of the IEP or LEP students are included, their results should not be reported separately.

With regard to sampling and weighting for nonparticipation, attention should be given to stratifying public schools within a given state by test scores on statewide exams.

Attention should be given to estimating the amount of systematic error caused by nonparticipation of schools and of students within schools. Attention should also be given to developing additional sources of data to adjust for school and student nonparticipation.

## *A Study of the Administration of the 1994 Trial State Assessment*

Liz Hartka
Jin-Ying Yu
Don McLaughlin
*American Institutes for Research*

### *Introduction*

As part of the NAE evaluation of the TSA, the American Institutes for Research (AIR) has carried out studies of the administration of the assessment in 1990, 1992, and 1994. In general, each of these studies has examined the extent to which assessment procedures are reliably implemented in the field and whether variations in implementation affect the quality of the data obtained.

The 1990 study concluded that the assessment of eighth-grade mathematics was carefully planned and administered, and that training was adequate for the assessment. However, recommendations for future refinements of assessment procedures were made. The 1992 TSA added a new level (grade four), a new content area (reading), and more participating states to the assessment. Classroom observations of the 1992 TSA noted only minor problems with assessment procedures; these included problems with the paced tape administration of the estimation items in the mathematics assessment. The two most common problems reported in a survey of

state assessment directors were the difficulty in procuring schools to participate and the time required of staff for administration.

In 1992 positive findings included the fact that most assessment administrators diligently followed the instructions for the assessment and completed their tasks with care. The survey of state testing directors indicated that respondents generally were satisfied with the conduct of the assessment in their states. Variance components analyses of student performance data demonstrated that there were no significant differences between scores in the TSA and the national comparison samples, and performance in monitored and unmonitored classrooms seemed to be about the same. Mean performance levels in assessment sessions (across all grades and subjects) were related to the following characteristics of the sessions: ability of the assessment administrator to complete the administration and related paper work independently, student cooperation, student attitude toward the assessment, and status of the assessment administrator. The report concluded with several recommendations, including recommendations that session monitoring be continued at a reduced level and that better administration and background data collection procedures be developed for the younger (grade-four) students.

## *Methods*

For the 1994 TSA, the administration study drew on three main sources of information.

1) ***Classroom observations conducted in February 1994.*** Classroom observations were designed to provide a first-hand look at the implementation of assessment procedures in the field and to give an indication of how well the TSA was administered. Panel staff visited 56 schools that were systematically sampled from several regions in the United States. Both monitored and unmonitored, and public and nonpublic sessions were observed. Schools in low income areas were over sampled because observers of the 1992 administration study reported seeing more implementation problems in these types of schools. AIR observers attended assessment administrator training sessions conducted by Westat in northern California, and also participated in an in-house training session that instructed them in the use of an observational protocol designed specifically for the study. In addition to filling out the observational protocol, observers provided written narratives describing their sessions.

2) ***Variance components analyses based on student performance scores and data on session-level irregularities***. Variance components between states, between sessions in the same state, and within sessions were calculated and tested to address the question of whether inconsistent administration conditions were associated with variations in student performance. The focal sessions, in which specific irregularities were observed, were identified using the information gathered from the monitoring and debriefing forms filled out by the Westat quality control monitors. Variance components analysis was also used to test the comparability of student performance in monitored and unmonitored sessions and in the national and state assessments.

3) ***A survey of state testing directors conducted in summer 1994.*** This mail survey was designed to determine the overall level of satisfaction that state testing directors had with the 1994 TSA. Questionnaires were mailed to the state testing directors in May, 1994, and responses were obtained from 37 participating states and territories and 10 nonparticipating jurisdictions, for an overall response rate of 87 percent. State testing directors from participating states were asked about the general conduct of the TSA in their states, nonparticipating states were asked why they did not join the TSA program, and all respondents were asked about both their expectations for the TSA program and their plans for future participation.

## *Results and Discussion*

### *1) How well was the 1994 TSA administered?*

Overall, results showed a reasonable level of compliance with the established protocol for conducting the assessment. Most of the 56 sessions observed by Panel staff proceeded well. Assessment administrators were careful to observe the procedures, and students were well behaved and worked hard on the assessment. In several cases however, there were major departures from procedures and the sessions did not proceed smoothly. Departures from the script occurred most often in two particular sections of the assessment: "Coding the Booklet Cover" and "Ending the Session." Fourth graders had the same problems with the student background questions that were witnessed in 1992. In particular, students were confused by questions about race/ethnicity, reduced price lunches, parent education, and other parent (especially step-parent) characteristics.

The 1994 TSA reading assessment included two 25-minute item blocks which were supposed to be strictly timed. Panel observers noted that several assessment administrators called out a 10-minute warning (i.e., to announce the remaining time for a given block) at the wrong time.

Sessions lasted an average of one-hour-and-39 minutes. Although most students were cooperative and orderly during the assessment, some did get tired towards the end. Students seemed generally to work longer and harder on the assessment when the assessment administrators encouraged them to do so, and were noticeably better behaved when the assessment administrator was the school principal, despite the fact that school principals were not always the best assessment administrators. Previous administration studies have, in fact, consistently found that higher student performance is associated with higher status administrators, such as principals, assistant principals, or visiting superintendents.

Several observers remarked that sessions started on time and proceeded more smoothly when the schools prepared in advance for the assessment (as they were instructed to do). Optimal situations were those in which assessment administrators and school administrators cooperated to ensure that students arrived on time and that sessions were not plagued by interruptions. Observers also felt that local support of the assessment tended to create a good environment for the sessions and encouraged

positive student attitudes. In addition, assessment administrators who prepared for the assessment did a much better job of conducting the sessions. Observers remarked that the performance of assessment administrators who had practiced the script, for example, was noticeably better. Discomfort with the assessment and unfamiliarity with procedures were viewed as detrimental to the administration by Panel observers.

In summary, the main problems with the administration of the assessment appeared to be that fourth graders had trouble understanding the general background questions, some assessment administrators seemed confused about how much help was permissible during the assessment, and some called out the 10-minute warning at the wrong time. In addition, many students had trouble coding their school identification numbers and home zip codes onto their booklet covers. These glitches could probably be ironed out in future assessments with additional training (with respect to the timers) or with additional paperwork (by having the assessment administrators fill in students' booklet covers for them).

### 2) *Were departures from scripted procedures associated with differences in student performance?*

Findings from the variance component analyses identified three types of factors that were associated with higher average student performance:

On-time completion of pre-assessment tasks (i.e., supplemental sampling and distributing Teacher Questionnaires);

Competent administration of the assessment and accurate reading of the script (especially the "Directions" and the general "Background" sections); and

Positive student attitudes and cooperation.

All of these results point to the desirability of reinforcing the preparation for assessment administrators, particularly the requirement that assessment administrators practice the script before they oversee the assessment. It is also apparently helpful to have staff at participating schools project a positive attitude about the assessment to the students. If staff treat the TSA as a valuable activity, students will tend to take it more seriously.

### 3) *Were there any differences in performance between the national and state assessments?*

Results of the variance components analysis of differences between the national and state samples showed that performance did not differ significantly between the two groups, although performance in the TSA sample was slightly higher. These results were consistent with results from the 1992 fourth-grade reading TSA, and indicate that assessment conditions in the state and national samples were probably similar and comparisons between them justified.

Within-session (i.e., within-school) variance components decreased slightly between 1992 and 1994. This reduction in within-session variability was accompanied by increases in both between-session (i.e., within-state) and between-state variance. This occurred in both state and national samples.

### 4) *Was student performance about the same in both monitored and unmonitored sessions?*

Differences between monitored and unmonitored sessions were statistically significant: monitored sessions averaged higher scores on the TSA than unmonitored sessions. When the 1994 session data were broken out by public versus nonpublic schools however, the significant performance differences between monitored and unmonitored sessions disappeared. A proportionately greater number of nonpublic schools (which tend to have higher-scoring students) appeared in the monitored category because nonpublic schools were monitored at a rate of 50 percent, as compared to a 25 percent rate in most states' public schools. Thus, when public and nonpublic schools were combined in the same analysis, the higher overall mean for monitored sessions appeared to be due primarily to the inclusion of proportionally greater numbers of nonpublic schools in this category.

### 5) *Was the experience with the 1994 TSA at the state level positive or negative?*

States that participated in the 1994 TSA indicated that, procedurally, the program went well in their states; in fact, overall levels of satisfaction with the 1994 implementation *increased* substantially from 1992. Testing directors were generally satisfied with the amount of time required for paperwork, training, and administration of the TSA. The main problem reported by participating states involved securing nonpublic school participation; over half (54 percent) reported problems with private school recruitment. On the other hand, the proportion of states reporting problems recruiting *public* schools decreased from a high of 48 percent in 1992 to 24 percent, which is approximately the same level reported in 1990. This may have been related to the fact that the 1994 TSA, like the 1990 TSA, only encompassed one subject at one grade; much larger numbers of students and schools were required for the 1992 TSA.

### 6) *What are the expectations at the state level regarding the future of the TSA program?*

The responses of the state testing directors reflected high levels of satisfaction with the conduct of the 1994 assessment. An examination of responses from earlier surveys of the testing directors showed that state satisfaction with the TSA program has been on the increase. However, many state testing directors reported problems with the inclusion of nonpublic schools. In response to questions about future participation in the state NAEP program, currently nonparticipating states continued to cite cost and burden to staff and schools as deterrents to joining the TSA program. State testing directors in currently participating states indicated that the ability to predict which content areas and grade levels would be included in the state NAEP (presumably more than one year in advance) would help to ensure their continued participation in the program. In addition, state testing directors in all states said that having the ability to choose among content areas and grade levels to tailor the TSA to the needs of their own state would provide a strong incentive to participate.

In summary, although state satisfaction with the TSA is on the increase, there is still uncertainty among states about the future of the state NAEP program. States continue to express concerns about the cost of the program and are unwilling to share a larger portion of the monetary burden. Finally, states seem to be keenly interested in having a state NAEP program that can be tailored to their own assessment needs.

The following recommendations are made:

If NAEP continues to collect general background information directly from fourth graders (as opposed to administering a parental questionnaire), assessment administrator training should place more emphasis on helping younger students through this section. Many assessment administrators seemed confused about what kinds of assistance they were permitted to offer students. Perhaps role playing during the training would help, especially for questions addressing race/ethnicity and parents in the home—issues that seem to be most confusing for fourth graders.

Assessment administrator training should continue to emphasize scrupulous adherence to the script and the rest of the procedures for the TSA. In addition, incentives should be given to encourage assessment administrators to practice reading the script and going over the procedures before the assessment. Westat should consider implementing either additional training or review sessions for assessment administrators who need extra help.

Regular training should emphasize the importance of keeping students on-task during the assessment. Students, particularly fourth graders, seem to need encouragement to check their work after completion of each item block.

Half of the states responding to AIR's survey of state testing directors reported problems securing nonpublic school participation in the 1994 TSA. The Panel understands that NCES is planning to have the nonpublic schools recruited by Westat during the 1996 state by state. This decision is to be commended.

Funding uncertainties and cost issues have been offered as reasons that preclude setting the full schedule with any certainty more than a year in advance, but some compromise may be necessary if consistent participation from the states is important to the success of the program. For example, having a "basic" schedule, whereby fourth-grade reading and mathematics and eighth-grade science are assessed according to a fixed schedule (which could be spread over four to six years), would be better than having no fixed schedule at all.

The issue of monitoring TSA sessions deserves continued vigilance by NAEP. The level of monitoring should not drop below the current proportion of 25 percent of sessions in continuing states, and the 50 percent level in new states.

# Public School Nonparticipation Study

Liz Hartka
Marianne Perie
Don McLaughlin
*American Institutes for Research*

## Introduction

The public school nonparticipation study was undertaken to investigate the association of higher state performance on the assessment with lower initial or before-substitution school participation rates. In fact, negative correlations between initial school participation rates and average state performance were observed for each of the TSAs: 1990, 1992, and 1994. This unsettling observation raised concerns that performance differences between participating and refusing schools may have been biasing the state NAEP results.

The investigation focused on the statistical relationship between public school nonparticipation and overall state performance on the 1994 TSA. It also examined factors (at the state or school level) associated with differential school participation rates in the TSA.

## Methods

Panel staff at AIR used auxiliary data to estimate the impact of nonresponding schools on state performance on the TSA. State assessment data for 11 states were secured for the purpose of estimating the performance of nonresponding schools on the TSA. Most correlations between state assessment scores and TSA scores ranged from approximately .6 to .8, indicating a moderately strong linear relationship between the two measures. Multiple regression models that included state assessment school means, as well as demographic information from the NCES Common Core of Data (CCD), were used to predict TSA scores for the refusing schools. The proportions of variance accounted for by the regression equations used to predict TSA scores ranged from about 50 percent to 80 percent, again indicating a good fit of the linear model to the data. One hazard of using the regression model, however, is the effect of "regression toward the mean" on predicted scores; this shrinkage in variability among the predicted scores ranged from close to 0 in one state to about 40 percent in another state.

## Results and Discussion

Variables for the analyses included predicted school mean TSA scores, actual school mean TSA scores (for participating schools), state assessment school means, and additional school information from the CCD. Weighted before- and after-substitution school samples were compared, state by state, using the predicted TSA school means as proxies for refusing schools in the before-substitution samples and actual TSA

means for all other schools. There were no statistically significant differences between these means for any of the states analyzed. A similar analysis was conducted using predicted TSA means for *all schools* (refusing, substitute, and cooperating) as the dependent variable. Again, there were no significant differences among the states analyzed.

A third weighted comparison was conducted using state assessment scores (not TSA scores) as the dependent variable for both before- and after-substitution samples. Once again, the means for this comparison were virtually identical, indicating very little (if any) bias due to substitution in the results.

Finally, direct comparisons (both weighted and unweighted) were made between predicted mean TSA performance for refusing versus substitute schools within each of the 11 states in the study. A significant difference was found in only one state, in Rhode Island, under the weighted comparison. The predicted performance of refusing and substitute schools on the TSA was approximately the same in each state, and there was little evidence that the substitution process had yielded flawed results. Refusing and substitute schools were also compared in terms of performance on their own state assessments. Again, only one significant difference, in Montana, was found between the samples in the weighted comparison.

Refusing and substitute schools were also compared with respect to a series of descriptive variables from the CCD data set. These comparisons indicated that refusing schools tended to be larger schools servicing student populations that were predominantly racial/ethnic minority group members.

## *Summary and Recommendations*

Based on the weight of the evidence, the investigator concluded that the effects of public school nonparticipation were negligible in the 11 states examined in this study. Results indicated that the impact of school substitutions on overall state performance and subsequent rank on the 1994 TSA was very small in the subsample of states examined. In the aggregate, no significant differences between before- and after-substitution samples were found with respect to either predicted TSA performance or actual performance on each state's own primary-grade reading assessment. However, a more thorough investigation of differences between refusing and substitute schools, on a state-by-state basis, may be in order. It is also possible that, in some of the states not examined, larger differences exist between before- and after-substitution samples. Furthermore, the difficult issue of the negative correlation between initial participation rate and state performance still needs to be resolved. The most obvious explanation—that lower-performing refusing schools were replaced with higher-performing substitute schools—does not appear to apply in the 11 states that were examined in this study. However, it is still not clear why the negative correlation continues to occur.

On the basis of these findings, the following recommendations were made:

> Comparisons of before- and after-substitution samples using state assessment data should be done periodically (perhaps every other cycle) to examine the effects of nonparticipation on the assessment. Researchers and consumers of

NAEP data at the state level should find these analyses both useful and informative.

More information regarding the *reasons* that schools refuse to participate should be collected by state coordinators during the TSA recruitment period.

Although previous research has shown that the school nonresponse adjustments applied by NAEP adequately eliminate much of the bias (due to nonresponse) in the data, the investigator notes that *the best method for eliminating nonresponse bias is to eliminate nonresponse.* In states where full school participation is problematic or impossible however, recruiters should continue to be encouraged to make full use of substitute schools.

## *Study of Exclusion and Assessability of Students with Disabilities in the 1994 Trial State Assessment of the National Assessment of Educational Progress*

Fran Stancavage
Don McLaughlin
Robert Vergun
Cathy Godlewski
Jill Allen
*American Institutes for Research*

### *Introduction*

In response to recent education trends, programs conducting large-scale assessments have been re-evaluating their guidelines and procedures for assessing students with disabilities. NAEP is among those programs that are attempting to include greater numbers of IEP students in their assessments, and it has recently made several changes to its exclusion procedures. For example, assessment accommodations were made available to samples of students who participated in the 1996 national mathematics and science assessments. In addition, assessment administrators from many of the schools in the national and state NAEP programs operated under revised inclusion guidelines in 1996.

These changes were informed, in part, by the findings from the Panel's study, reported here, that explored the assessability of excluded students with disabilities in the population sampled by NAEP. The study also examined the possible need for accommodations to include additional students, the implementation of the exclusion decision process in the 1994 state NAEP fourth-grade reading assessment, and the extent to which differences in the application of the exclusion guidelines between states affected the accuracy and fairness of state NAEP data.

## Methods

The sample consisted of students selected for the 1994 fourth-grade reading TSA who were identified by their local school personnel as having individualized education plans. Some of these students actually participated in the TSA while others were excluded on the basis of their disability. Site visitors met with the students to obtain measures of reading proficiency, and interviewed students' teachers and local NAEP assessment administrators to collect additional information about the students and the implementation of the exclusion process.

The investigators selected four states that represented both high and low identification rates and high and low conditional exclusion rates of IEP students. The final sample included 123 schools distributed across these four study states. Among the 444 eligible students in the study schools, student and/or teacher data were collected for 416.

Experienced teachers or researchers who had worked with fourth-grade students and/or students with disabilities were hired and trained to collect the data. During their site visits, they obtained an independent assessment of each student's reading ability using an adaptive protocol which included a brief rapport building activity, the WJ BRC, and an abridged, and relatively easy, item block from the 1992 TSA. The site visitors also conducted structured interviews with teachers or staff persons familiar with each student's disability status, level of functioning, and school program. This teacher interview included several questions about testing accommodations as well as questions about NAEP exclusion decision procedures and guidelines. Similarly, the interview with the NAEP assessment administrator included questions about the respondent's general impression of the NAEP exclusion decision guidelines and procedures.

## Results and Conclusions

### *Assessability*

The investigators first addressed the question of assessability by examining the relationship between the WJ BRC scores and the NAEP plausible values of included IEP students in the sample. The results showed that fourth-grade students with grade-equivalent WJ BRC scores at or above 2.1 could perform meaningfully on NAEP. Below 2.1, NAEP plausible values no longer appeared to capture any of the observed variation in the proficiency distribution. The researchers cautioned, however, that the 2.1 cutpoint falls close to the bottom of the proficiency range for which comparative data were available. Hence, the 2.1 criterion should be interpreted as approximate rather than precise. Further analyses revealed that fully 83 percent of all the students in the study sample, both included and excluded, had scores at or above 2.1 and therefore met the criterion for assessability. The assessable group included 70 percent of the excluded students and 93 percent of the students who had participated in the TSA.

The investigators obtained a second perspective on the assessability of IEP students by reviewing their performance on the full individual student assessment administered for the study. While the analyses indicated that the reading ability of the majority of

students with disabilities can be assessed by NAEP, the assessment is clearly not pitched at a level that can provide a fine-grained measure of reading for these students, many of whom are reading a grade or more below grade level.

***Accommodations***

To address the second research question, concerning assessment accommodations, site visitors asked teacher respondents whether their students should have taken the assessment as given, should have taken it with accommodations, or more appropriately should have been excluded. Teachers favored accommodations for more than half of their students with disabilities, including nearly two-thirds of those who had been included in the 1994 TSA and over 40 percent of those who had been excluded. For nearly all the remaining students, teachers provided recommendations identical to those given at the time of the TSA assessment.

In response to an inquiry asking teachers about the types of accommodations they would recommend, they stated most frequently that learning disabled students (who accounted for over two-thirds of the study sample and a similar proportion of the full IEP population sampled by NAEP) should be given additional time and/or shorter tests. Other frequently mentioned accommodations for learning disabled students included oral reading of directions, small group or individual administration, and oral response. Students with orthopedic or sensory disabilities, who might require more specialized or technologically sophisticated accommodations, accounted for a very small fraction of the study sample and only 2 percent of the IEP population sampled by NAEP.

***Exclusion Process***

Both the teacher and assessment administrator respondents frequently reported being involved in the exclusion decisions, often in consultation with one another or with other members of the IEP team. Forty-five percent of the teacher respondents reported that they had seen copies of the official NAEP exclusion guidelines when they made their decisions, 39 percent indicated that they did not see them, and 16 percent did not remember. All of the assessment administrators would have seen these guidelines since they were included in their NAEP assessment training manual. Most of the teacher respondents (91 percent) and assessment administrators (81 percent) reported finding the guidelines clear.

The most frequent reasons for not finding the guidelines clear involved trying to apply the "percent time mainstreamed" criteria in schools that served their students with disabilities in the regular classroom setting, and being uncertain how to operationalize the phrase "participate meaningfully in the assessment." When given a list of pre-specified factors that may have influenced exclusion decisions, respondents who had participated in the exclusion process most frequently reported that the student's reading level was a primary factor. The second most common factor cited as influencing the decision to exclude was that the IEP specified that the student should/should not participate. Overall, while it appeared that the teachers and assessment administrators were abiding with the spirit of the NAEP guidelines, they tended to have a conservative view of the level of functioning that was necessary for students with disabilities to participate "meaningfully" in the assessment.

***Comparability Between States***

Lastly, the investigators explored the comparability of exclusions across states. Results showed that all four states tended to exclude poorer readers more than better readers, although there was substantial overlap in the reading score distributions of included and excluded students in each state. When the exclusion cutpoint was defined as the reading grade level at which the probability of exclusion reaches 50 percent, results showed that the states differed significantly in their criterion for exclusion of IEP students from the 1994 TSA.

## *Recommendations*

Based on the findings described above, the investigators offered the following recommendations:

NAEP should continue efforts to encourage greater participation of students with disabilities on the current assessment. In addition, the results for students with disabilities assessed under standard conditions should be aggregated with results for all other students in producing the overall and subgroup achievement estimates normally reported for the nation and the states. NAEP should also work to develop assessments that can measure accurately over a broader range of student achievement levels and thereby provide better estimates at both ends of the achievement distribution.

NAEP should continue to offer accommodations to students with disabilities in order to increase participation. Furthermore, NAEP should develop guidelines for the use of accommodations, continue to perform research into the impact of accommodations, and consider standardizing the choice of accommodations to the extent feasible.

NAEP should revise the exclusion guidelines to specify more concrete criteria for inclusion. Moreover, NAEP should consider the advisability of providing its own estimation of the functional reading level required for meaningful participation.

# *Study of Exclusion and Assessability of Limited English Proficiency Students in the 1994 Trial State Assessment of the National Assessment of Educational Progress*

Fran Stancavage
Jill Allen
Cathy Godlewski

## *Introduction*

*Goals 2000: Educate America Act* calls for academic standards and assessments that are meaningful, challenging, and appropriate for all students. Representing a recent trend in education, the passage of this law has prompted greater efforts to include LEP students in national and state assessments. In conjunction with NAEP contractors, NCES and NAGB have demonstrated their commitment to increasing inclusion of LEP students and have begun to implement a number of changes in the 1996 NAEP assessments, informed in part by findings of the Panel's study, described here. This study investigated the assessability of the currently excluded LEP students on the 1994 NAEP reading assessment, the accommodations or adaptions that would be needed to include additional LEP students, and the implementation of the exclusion decision process for LEP students in the 1994 TSA.

## *Methods*

The investigators of this study selected a purposive sample from three states which had high proportions of LEP students. Within these states, Spanish bilingual site visitors collected data for 254 students at 65 schools. All of these students had been selected for the 1994 fourth-grade reading TSA and identified as LEP by their local school personnel. Some of the students actually participated in the TSA while others were excluded. The exclusion guidelines in effect in 1994 stated that students could be excluded if they were native speakers of a language other than English, enrolled in an English-speaking school (not including a bilingual education program) for less than two years, and judged incapable of taking part in the assessment.

Site visitors obtained an independent measure of the students' ability to read English. Because experts on second language learners consulted by the investigators could not agree on a standardized measure of English language proficiency for the LEP students, site visitors used a brief second-grade story to test for comprehension and oral reading ability. If the student performed reasonably well on the oral reading and retelling task, the site visitor then administered an abridged NAEP item block from the 1992 TSA. Site visitors also interviewed students' teachers and local NAEP assessment administrators about the students' academic programs and level of functioning, recommended testing accommodations and adaptations, and local implementation of exclusion decision procedures and guidelines.

All of the sampled students who participated in the 1994 TSA and more than 90 percent of the excluded students performed well enough on the second-grade story to progress to the NAEP item block. Despite the fact that the NAEP item block was considerably more difficult than the second-grade story, 75 percent of all sampled students answered one or more of the multiple-choice items correctly. Among those who answered at least one item correctly, 86 percent were included students and 57 percent were excluded students. The constructed-response items proved to be more difficult for the LEP students. Thirty-nine percent of the included students and only 16 percent of the excluded students answered at least one constructed-response item correctly. Overall, the evidence demonstrated that a high proportion of LEP students would have been assessable on the current NAEP instrument.

Teacher respondents were briefed on the broad characteristics of the NAEP reading assessment and shown a sample item block. They were then asked, in light of the length and the format of the assessment, whether their LEP students could have taken the assessment, or could have taken it with certain accommodations or adaptations. The teachers frequently recommended that these LEP students be included with accommodations or adaptations, or excluded altogether from TSA participation. Specifically, the teachers recommended exclusion for 26 percent of all the sampled students, including 10 percent of the students who had participated in the TSA and 51 percent of the students who had not. They also recommended that 45 percent of both included and excluded students should be included in the TSA, but with accommodations or adaptations.

At the end of the interview, teachers were presented with choices in four different categories of accommodations/adaptations (presentation format, response format, setting of test, and timing of test) and asked to identify which of these their students would need to participate in the TSA. Teachers indicated most frequently that LEP students should be given extended testing time and/or a shorter version of the test (82 percent). For 75 percent of the LEP students, teachers recommended that NAEP provide presentation format accommodations, namely, pictures or diagrams, with or without instructions in the student's native language and/or out-of-grade testing. In addition, teachers reported that 52 percent of LEP students should be tested alone, in small groups, or in a special language/education class. Another 33 percent of LEP students were described as needing assistance and interpretation with their response.

Thirty-five percent of the teacher respondents reported that they saw a copy of the official NAEP exclusion guidelines when they made their decision, 29 percent did not see the guidelines, and 35 percent did not remember. All of the assessment administrators would have seen the guidelines since they were included in their NAEP assessment training manual. The majority of the teacher respondents (83 percent) and assessment administrators (76 percent) reported that the guidelines were clear.

Generally, the respondents indicated that the inclusion and exclusion decisions were reached in a group process, with several individuals participating. The most commonly reported reasons for including or excluding a student from the TSA were that the student "reads/doesn't read well enough" or is "reading/not reading at grade

level" (32 percent). Other reasons selected from the list of suggested factors included the "TSA is appropriate/not appropriate to student's instructional program" (26 percent), the "student was included/not included in the state assessment" (17 percent), and the student "understands/doesn't understand oral English well enough" (12 percent).

The investigators of this study performed a multivariate logistic regression to explore the effects of other factors on the decision to exclude. The results showed that the percentage of time per week spent in a special language (bilingual) program and the exclusion from state, district, or other grade-level standardized tests were *positively* associated with exclusion. In other words, LEP students who spent more time in a special language program or who were excluded from other tests were more likely to be excluded from the TSA than students who were not. In contrast, the number of years spent in a special language program and the teacher's (higher) estimation of the student's functional grade level for writing English were *negatively* associated with exclusion. Restated, students who spent more years in a special language program and whose teachers gave higher estimates of their functional grade level for writing were less likely to be excluded.

## *Recommendations*

The investigators concluded with the following recommendations:

NAEP should continue to encourage greater participation of LEP students in the current assessment.

NAEP should continue efforts to identify appropriate adaptations or accommodations that would permit the inclusion of even larger proportions of LEP students in the assessments.

# *The 1994 Reading Anomaly: Report to The National Academy of Education on the Drop in the National Assessment of Educational Progress Main Assessment (Short-Term Trend) Scores*

Larry V. Hedges
*University of Chicago*

Richard L. Venezky
*University of Delaware*

## *Introduction*

In preparation for the release of the 1994 reading results, NCES requested that the NAE Panel on the evaluation of the TSA examine the drop in reading scores at grade 12.[1] In this report, the authors have reviewed a number of hypotheses that might have accounted for the observed drop in reading scores, including true change in achievement, a drop in motivation, a change in the effective sample populations, equating bias, effects of particular items or item types, and booklet or block effects.[2] Based on the evidence available, the authors concluded that the observed decline from 1992 scores could not, with a high degree of assurance, be attributed only to a real decline in student reading ability. Although they were not able to isolate a set of variables that accounted for all of the score difference, they identified certain variables that might have caused a substantial portion of the difference observed.

## *The Score Decline*

The decline in short-term trend scores from 1992 through 1994 occurred at grade levels 4 and 12, although only the decline at grade 12 was significant. The size of the composite score decline at grade 12 was 4.5 points on a scale that extends from 0 to 500, with a standard deviation of 37. The decline was therefore 0.12 standard deviation units, or, a difference from 1992 of slightly less than one item correct per student. Although this change does not seem large in absolute terms, it is larger than

[1] This report was prepared in April, 1995 and is based on the data available at that time. A postscript has been added to update the conclusions with information made available since that time.

[2] The best evidence for a true change in ability would be corroboration from the results of the 1994 long-term trend assessment in reading. However, only preliminary results for the long-term assessment were available at the time of the authors' analyses. Moreover, the investigation was constrained by the fact that only two data points (1992 and 1994) were available for establishing a short-term trend. Following completion of this report, the authors received the results of the long-term trend survey for grade 11 as well as notification of an error in the scaling program used by ETS for computing the 1992 and 1994 NAEP reading and mathematics scale scores. Accordingly, they append a postscript stating that the long-term trend results appear similar to the main assessment, although two differences are noteworthy: 1) none of the changes in the long-term trend are significant and 2) scores of black students increased rather than decreased (as in the main assessment) from 1992 through 1994. In reference to the scoring error, the authors state that the discovery of such an error should lead to a strengthening of both internal and external review of NAEP results before they are reported.

any other reported for NAEP reading trend assessments since 1980 (except for the 1986 assessment, which, after extensive investigation, was found to have been contaminated by methodological changes).

The 1994 decline affected all subpopulations normally reported. However, it was greatest for those scoring below the 50th percentile, for those whose parents had only a high school education or less, for males, and for black and Hispanic students. At the 12th-grade level, white, black, and Hispanic students all scored significantly lower in 1994 than they had in 1992. The decline was greatest for the "reading to perform a task" scale, although all three scales displayed lower scores.

Although the largest declines occurred among those who generally perform poorly, statistically significant declines also occurred among some of the subpopulations that usually perform at the highest levels. For example, males who said that they read a story, novel, or newspaper every day, or who rated themselves as very good readers also achieved lower scores in 1994 than in 1992. In contrast, female scores in these categories neither changed significantly nor displayed a downward trend.

An additional characteristic of the differences between 1992 and 1994 results was an 11 percent increase in the standard deviation of the 1994 results, most of which appeared to be a result of an increase in low-performing students. That is, the lower tail of the 1994 score distribution extended farther toward the zero point than did the tail of the 1992 distribution.

## *Test for a True Decline in Performance*

To test the relation between the decline in NAEP scores and a true decline in performance, the authors pursued two hypotheses: that the decline resulted from a decline in true ability, and that the decline resulted from a drop in test taking motivation.

Long-term trend data could have corroborated a true decline in ability. This data was not available at the time of the authors' analyses however. With the cooperation of the Council of Chief State School Officers (CCSSO), NCES, and various other agencies and individuals, the authors therefore investigated the available non-NAEP data on 12th-grade reading performance. They found no evidence of a general, nationwide decline in reading or language arts performance. Individual instances of state or district testing, which are highly limited at the 12th-grade level, showed more declines than increases, but the authors found no national downward trend. The failure to find strong convergent evidence for a true decline in ability does not falsify the hypothesis that such a decline was responsible for the observed decline in NAEP scores; it only makes it less likely to be true.

The authors found no evidence suggesting that the observed NAEP decline was due to a change in test taking motivation. They did note however that various student practices related to literacy and schooling had also declined from 1992 through 1994, according to the NAEP self-report data. These practices included the percentage of students reading daily or weekly for fun, reading 15 or more pages per day, and spending one or more

hour per day on homework. Although these variables tend to correlate positively with NAEP reading scores when other variables are not controlled, the relationships are found only for individuals within an assessment, not for group means across assessments.

## *Tests for Changes in the Effective Sample Population*

The authors did find evidence suggesting that the sampled populations in 1992 and 1994 may have been nonequivalent. They found, for example, that responses to the school background questionnaires showed that, on average, the schools in the 1992 sample required more semesters of English, math, and science for graduation than did those included in the 1994 sample. The two samples also differed in the percentage of schools which reported 1) no students with subsidized lunches, 2) more than 10 percent student enrollment in remedial reading, 3) no English as a second language (ESL) students, and 4) more than 10 percent student absence rates on an average day. Although few of these differences were statistically significant, the authors reported that these factors and others, such as the numbers of counselors and psychologists working in the schools, appeared to indicate a less academically demanding and a more socially, economically, and academically challenged pool of schools in 1994.

Another indicator of a shift in the sample was found in the score changes of students who had the highest literacy practices and were in schooling categories that normally achieve high-scale scores. For example, scores declined for students who 1) read 20 or more pages each day and whose parents had graduated from college, 2) read five or more books outside of school during the past month and attended a private or Catholic school, and 3) read five or more books outside class and attended a school that ranked academically in the top one-third of U.S. schools. Although none of these differences were statistically significant, they suggest that, even among those in the 1994 population from whom one would expect the highest scores, scores declined.

## *Equating from 1992 through 1994*

Problems in equating the 1992 and 1994 assessment scales represented another possible factor in the decline. All NAEP assessments have a mixture of multiple-choice, short constructed-response, and extended constructed-response items. Although there were many common constructed-response items on the 1994 and 1992 assessments, changes in the performance of the constructed-response items meant that many could not be used for equating. Consequently, a disproportionate number of the items used in equating were multiple choice. Because the distribution of items used for equating differed from the distribution of items used in the operational scales, the possibility arises that the scales were equated based on a weighting of factors different from those used in the operational scales.

## *Item, Booklet, and Block Effects*

The authors found no evidence for item type effects when they looked at the items common in the 1992 and 1994 assessments; nor did they find evidence to suggest that the decline was due to booklet or block effects.

## *Summary and Recommendations*

In summary, the authors found no single methodological factor that accounted for the decline, but instead found a number of factors that, taken together, plausibly could have created the decline. These included a potential shift in the effective sample population, under-representation of constructed-response items in the common pool of equating items, and errors in the estimates of the mean introduced by the equating algorithm. The evidence available for the shift in the population sampled would have been consistent with a decline in 1994, but the evidence for the equating items and the equating algorithm was nondirectional.

The authors concluded with four recommendations:

> "When measuring change, don't change the measure," as stated by Al Beaton in the first report on the 1986 reading anomaly.
>
> Analyze related NAEP data sets that might corroborate or contradict findings before releasing assessment results to the public.
>
> Report assessment outcomes for subpopulations that are relevant for policy formation. Reporting by groupings according to, for example, destination after high school (e.g., college bound, work bound), language (English, ESL), and the extremes of the performance distribution are important for policy and also for interpretation of results.
>
> Strengthen the external review of NAEP and include external indicators that could provide a composite picture of the progress of schooling in America.

# *Reporting the 1994 Reading Results by Achievement Levels*

George W. Bohrnstedt
Evelyn F. Hawkins
*American Institutes for Research*

## *Introduction*

In December, 1989, NAGB initiated a process to set performance standards for NAEP. These standards, which NAGB called achievement levels, came under considerable examination and engendered a national discussion regarding their validity and utility for reporting the state of educational progress in the nation. The controversy has focused on both the achievement-level cutscores themselves and on the modified Angoff method used to set them.

Because of the national interest in education standards, and because the achievement levels were both a new and controversial part of the NAEP program, NCES asked the Panel to conduct an independent evaluation of the 1992 achievement levels in reading and mathematics. The Panel gave direct attention to the adequacy of the achievement-levels setting process itself, and to the reliability of the inferences about levels of student performance drawn from these levels. The evaluation involved a series of internal validity and external comparison studies.

This paper is a review of the issues involved in the controversy. It describes some of the evidence from the NAE Panel's evaluation of the 1992 achievement levels and analyzes the data from NAGB's most recent attempt to validate the resulting achievement levels: the revisitation study of the 1992 reading achievement levels. Additionally, the paper considers the 1994 U.S. history and world geography achievement levels in an effort to obtain a more comprehensive picture of the functioning of the achievement levels. The paper concludes with recommendations regarding the continuing use of achievement levels for reporting NAEP results.

## *Review of Process of Setting Achievement Levels*

NAGB selected a modified Angoff procedure to set the achievement levels for NAEP. The Governing Board first adopted generic definitions of basic, proficient, and advanced performance levels, then used them to develop more specific descriptions of the three performance levels in each NAEP subject area and at each grade assessed: 4, 8, and 12. Working from these subject-specific descriptions, expert judges first constructed mental pictures of students who would minimally meet the implied performance criteria for each achievement level. The judges then evaluated individual NAEP items and estimated the percentages of these hypothetical, marginally qualified students who they believed would answer each item correctly. The item-by-item

judgments were converted to the NAEP scale, then averaged across both items and judges to arrive at the final NAEP scale cutscores. The cutscores determined the lower boundaries of basic, proficient, and advanced performance at each of the grade levels.

***Findings from the Internal Validity Studies***

Evidence from the standard-setting sessions indicated that item characteristics that should have been irrelevant were actually strongly affecting the cutscores set by the judges. For example, cutscores set using dichotomous versus partial-credit (extended-response) items for the same grade and achievement level differed by as much as, or more than, the cutscores set for adjacent grades at the same achievement level using dichotomous items only. This was evident even though the level-setting process adjusted for differences in item difficulty. The Panel found similar internal inconsistencies for multiple-choice versus short answer questions and easy versus more difficult questions. The Panel concluded that standard-setting judges were unable to maintain a consistent view of their conceptions of basic, proficient, and advanced performance, and that, in fact, the judgements required were virtually impossible to make. The Panel concluded that the modified Angoff procedure used by NAGB had seriously deficiencies.

***Findings from the External Comparison Studies***

At grades 4, 8, and 12, the Panel found evidence indicating that more students were performing at higher achievement levels than the NAEP achievement-level cutscores identified. In three of four studies of fourth and eighth graders conducted on behalf of the Panel, teachers' ratings and individual assessments administered by researchers repeatedly found more students performing at the advanced, proficient, and basic levels than were classified by the achievement-level cutscores. Similarly, for grade 12, when data from the SAT and AP examinations were compared to NAEP results, higher percentages of students scored at what were considered advanced levels on the Verbal and Mathematics SAT and the AP examinations than were identified by NAEP.

Based on the findings of its evaluation of the 1992 achievement levels, the Panel recommended that NAGB discontinue the use of the Angoff method and urged NCES and NAGB not to report the 1992 NAEP results using the achievement levels.

## *The Revisitation of the 1992 Achievement Levels*

NAGB nevertheless decided to use the achievement levels for reporting the 1992 results. However, the Governing Board agreed to conduct an additional study of the 1992 reading achievement levels, the "revisitation study." Unfortunately, the design of the study addressed only the Panel's identification of a discrepancy between what students know and can do at the various levels and how the achievement levels are described. The study failed to address both the possibility that the cutscores were set too high and the issue of internal inconsistencies in judges' ratings by item type.

The revisitation study used two procedures to examine the possibility of a disjunction between the achievement levels actually used by NAGB and the descriptions implied

by items at or above the cutscores. In the first procedure, the judgmental item-categorization procedure, a panel of experts was asked to categorize the NAEP items into achievement levels based on the achievement-levels descriptions, then to assess whether, in general, the items adequately corresponded with the knowledge and skills described in the descriptions. In the second procedure, the item difficulty-categorization procedure, another panel was asked to examine how well students performing at a given achievement level performed on items in the assessment, then to assess the accuracy with which the items these students could or could not answer correctly reflected the knowledge and skills that the achievement-level description stipulated. The experts of both panels concluded that, at all three grade levels, students scoring at the given achievement levels knew and were generally able to do what the achievement-levels descriptions designated as the requisite knowledge and skills. They therefore supported the use of the 1992 achievement-level descriptions for reporting the 1994 NAEP reading assessment results.

The NAGB contractor for the revisitation study also conducted a series of statistical analyses on the data obtained in the study. Although the results of these analyses were intended to align with the conclusions of the expert advisors, the data, described in this paper, proved to provide less than compelling evidence in support of these conclusions. For example, as noted, experts of one of the panels were asked to categorize items by achievement level, according to the stipulated level of knowledge and skills required to get the item correct. The actual percentage of students getting each item correct was computed for each item that the judges had classified in each achievement-level category, and the percentages correct were averaged across the items in the category. These average percentages correct were then examined across achievement levels for each grade level, and were found to follow expected patterns: that is, the average percentage was higher for items categorized at the basic level than for those categorized at the proficient level, and was higher for those categorized at the proficient level than it was for items categorized at the advanced level. This held true for all grade levels. However, the percentages correct for items within each achievement level and for each grade level varied tremendously, thereby weakening the evidence provided by the simple average percentages.

The contractor also conducted rather complex analyses of "hit rates," in which students were classified on the basis of their NAEP scores, and items were classified based on the judgments of the expert panelists. Presumably, students who performed at a designated achievement level should have responded correctly to items categorized at, or below, that level, but not to items classified at a higher level. When NAEP results confirmed this pattern for a given item, the match was labeled a "hit." When the results disconfirmed the expected pattern, the mismatch was labeled a "miss." The hit rate was the number of hits divided by the total number of items at each respective grade level.

The contractor performed nine analyses of hit rates (three achievement levels by three grade levels) and found hit rates from 48 percent to 97 percent with a median hit rate of 81 percent. The expected hit rates on the basis of chance alone, however, ranged from 44 percent to 93 percent with a median of 72 percent. Thus, the number of hits was not far above what one would expect to observe on the basis of chance.

The review of the revisitation study demonstrated its failure to address the Panel's concern about whether or not the cutscores for the achievement levels were set too high. The study was not designed to address the Panel's conclusion that the modified

Angoff procedure had serious internal inconsistencies, and the data resulting from the study on the consistency between the achievement levels and the descriptions are less than convincing.

### *Continuing Issues regarding the Achievement-Levels Setting Procedure*

The achievement-levels setting process for the 1994 U.S. history and world geography NAEP was examined in order to obtain a more complete context for evaluating the reliability of the item-by-item achievement-levels setting procedure used by NAGB. Despite modifications made to the standard-setting procedures for the 1994 U.S. history and world geography achievement-level cutscores, the internal inconsistencies (noted above) remained, albeit at a somewhat diminished level. Moreover, a comparison between results for U.S. history and world geography raised additional questions of external validity for the achievement levels. The results showed that only 12 percent of 12th-grade students were classified as at or above the proficient level in U.S. history, compared to 29 percent in world geography, despite the fact that virtually every high school student in the nation takes U.S. history and very few take world geography. In addition, a comparison of the results of the 1994 U.S. history NAEP with those of the U. S. history AP examination showed more students performing at advanced levels on the AP examination than were identified by NAEP. These findings provide additional evidence to show that the Governing Board's achievement levels are identifying too few 12th-grade students performing at the advanced level.

### *Summary and Recommendations*

Because of fundamental problems with internal inconsistencies for the 1990 and 1992 mathematics and the 1992 reading achievement levels, continuing internal consistency problems with the 1994 history and world geography achievement levels, and apparent invalidities in the achievement levels when measured against external evidence of student achievement, this paper concludes in support of the Panel's belief that item-by-item methods for setting performance standards are fundamentally flawed as applied to student assessments such as NAEP. It recommends that NAGB explore alternatives to the modified Angoff procedure for setting achievement-level cutscores and that new cutscores obtained through alternative methods be empirically evaluated before they are made operational.

# Impact of the 1992 Trial State Assessment

Evelyn Hawkins
*American Institutes for Research*

## Introduction

As part of its evaluation of the NAEP TSA, the Panel conducted studies to examine the impact of the reporting of the 1990 TSA results on state education systems. The study reported here continues the examination of the impact of the TSA program, focusing on the 1992 TSA reports. The study attempted to answer two major research questions:

> What evidence exists that suggests that the 1992 TSA reports had an impact on state instructional or assessment programs in mathematics or reading?
>
> What aspects of the 1992 TSA reporting did states find useful, and how have the reports been used?

## Methodology

The assessment directors and the mathematics and reading specialists in each of the states and the District of Columbia (including both participating and nonparticipating states) were targeted for telephone surveys. The first round of surveys, conducted shortly after the release of the 1992 mathematics reports, began in June, 1993 and were concluded by September, 1993. Of the 102 planned surveys of assessment directors and mathematics curriculum specialists, 100 were completed.

The 1992 reading reports were released in September, 1993, and the second round of surveys was conducted between February and April, 1994. Ninety-five of the 102 assessment directors and reading specialists were surveyed.

## Research Findings

### *Impact on State Education System*

**Mathematics.** At the time of the first round of surveys, all but one of the 51 states[1] indicated that changes were being made in their mathematics instructional programs, and 47 of the 51 states indicated that changes were being made in their state mathematics assessment programs. Although many

[1] The District of Columbia will be referred to as a state to simplify the discussions.

of these reform efforts were already underway prior to the release of the 1992 TSA mathematics results, respondents indicated that the TSA influenced these instructional and assessment changes, primarily by validating and supporting the direction of the changes.

The 1992 TSA in eighth-grade mathematics was the second TSA in which this subject and grade had been administered, and it therefore provided trend data for the first time at the state level. Of those who were familiar with events following each of the two TSAs, the majority indicated that the 1990 TSA had had a greater impact on changes in the state mathematics programs than the 1992 TSA. Many of the respondents also indicated that impact could be best understood as deriving from the two TSAs in combination rather than simply one or the other.

**Reading**. During the period of the first two TSAs, reform efforts in reading instruction and assessment were somewhat less frequent than in mathematics. Of the 42 states that had participated in the 1992 TSA, 36 were making changes in their reading instructional programs (versus 41 in mathematics). However, among those that were making changes, a slightly larger percentage indicated that the 1992 TSA had influenced these changes. Similarly, in reading assessment programs, changes were being made in 32 of the participating states (versus 39 in mathematics), and a majority of these states (66 percent) indicated that the TSA had had an influence. The 1992 reading NAEP TSA may have had a greater influence than the 1992 mathematics TSA because it was the first TSA in reading.

**General Impact.** Many of the respondents in the participating states indicated that the 1992 TSA had had an overall positive impact on education in their states; another sizable proportion, however, saw the impact as too limited to characterize as positive or negative. None evaluated the overall impact as clearly negative.

***Use of NAEP TSA Reports***

In order to learn about how the TSA reports were used, respondents were asked about the actions their states took at the time of the release of the 1992 TSA results, the discussions regarding the TSA results that were held among different constituent groups, the ways in which the TSA reports were used, and the usefulness of the different reports and the various reporting formats. Most states reported having taken similar actions at the time of the release, specifically issuances of press releases and briefings with state-level staff. Most of the ensuing discussions regarding the results were held within the state departments of education and were fairly general in nature, concentrating largely on their own state's overall performance. Specific proposals to initiate reform were reported to have arisen in only a few states during these discussions.

The 1992 TSA results were published in various documents using a variety of graphical and numerical presentations. Respondents were asked about the ways in which each of these reports were used and about the presentation formats that were found to be particularly useful. All of the states indicated that they were using their own state reports. Many of them also reported using the executive summary of the

report card (presenting results for the nation and each participating state), the report card itself, and the data almanacs; the latter provided a tremendous amount of information for the interested reader but were not suitable for general dissemination. The reports were used for internal state department of education planning—both as a reference for background information and as a guide for improving instructional and assessment methods. Some of the information from the reports was also used for teacher in-service training. None of the states indicated that they had used the 1992 TSA reports to promote or support specific legislative actions.

After the first NCES reports of the 1990 mathematics NAEP TSA, NAGB issued its own report of the mathematics results, using achievement levels. The mathematics achievement levels were subsequently revised (and new levels set for reading), and reporting by achievement levels was incorporated into the main 1992 mathematics and reading NAEP reports. In the Panel's surveys, assessment directors and mathematics specialists were asked to compare the utility of the achievement levels with the anchor levels (descriptions of knowledge and skills associated with various points on the NAEP scale, but not tied to a performance standard) that had been used in the main 1990 reports. Respondents overwhelmingly found achievement levels more useful than the anchor levels, indicating that achievement levels were easier to understand, had more relevance for communicating the state of education, and therefore would be more likely to impact education policy. Reading specialists were asked whether they liked the general strategy of reporting performance using achievement levels. Out of the 37 responding reading specialists, 30, or 81 percent, answered in the affirmative.

Finally, in response to the different graphical presentations, respondents generally found the following formats to be useful: the customized maps comparing performance in the target state with performance in each of the other participating states, the quintile tables indicating where the state's performance for particular subgroups fell in relation to other states, and the map portraying the 1990 to 1992 change in eighth-grade mathematics performance for each participating state. The two map formats were found to be particularly useful in that they were easy to interpret and user friendly, had good layouts, and were suitable for use with general audiences and for oral presentations. General suggestions for improving the utility of the reports included using more graphic presentation instead of tables, highlighting differences and main effects with color, and streamlining the data presentation so users would not be overwhelmed with so much data.

## *Conclusions*

After the release of the 1992 TSA reports, most of the state assessment directors and subject area specialists from participating states felt that the NAEP TSA had had a positive impact on their mathematics and reading programs, albeit a relatively minor one. In most cases changes in state instructional and assessment programs were in progress, and the 1992 TSA results served primarily to support changes already being made. Further, the TSA results provoked discussions in various constituent groups and were particularly useful for teacher in-service training. The introduction of various new graphics and of achievement levels as a means of reporting results were well received.

The majority of respondents in all of the groups felt that the NAEP TSAs were a worthwhile endeavor. Those who answered positively explained that the TSA's value derived from the comparison it allowed between the states and the nation, and from the impetus it provides for change. Those who indicated that it was not, or only partly, worthwhile indicated that, in an era of tight budgets and competing priorities, the TSA was of questionable worth because it was not specifically linked to state curricular goals and could not provide results below the state level.

## *Perspectives on the Impact of the Trial State Assessments: State Assessment Directors, State Mathematics Specialists, and State Reading Specialists*

Liz Hartka
Fran Stancavage
*American Institutes for Research*

### *Introduction*

Throughout its evaluation of the TSAs, the Panel has collected information on the states' perceptions of impact and utility. The final round of data was collected during December, 1995 and January, 1996, nearly two years after the conduct of the last TSA in fourth-grade reading. At the time of the data collection, a summary of the 1994 state and national reading results had been available for about nine months, but the more comprehensive reading reports, including the individual state reports, had not yet been released.

Results of the 1995 data collection were generally consistent with trends observed in the Panel's earlier studies of state impact. That is, most participating states valued the state NAEP assessments as a model for good assessment practices and as an independent measure of student achievement against which they could evaluate the results from their own state assessments. The two factors most frequently cited as limiting the utility of the TSAs for the states were the long lag time to reporting—particularly salient in light of the extremely protracted reporting schedule for the 1994 assessment—and the lack of local or district results.

### *Methods*

Brief mail surveys were sent to assessment directors and mathematics and reading specialists in all 50 states and the District of Columbia. Questions focused on the overall impact of the TSA program since its inception in 1990, and respondents were asked specifically about recent or ongoing changes in reading and mathematics education practices, the impact of the TSAs on such changes, and the problems limiting the utility of the TSAs for the states. In addition, case studies covering most of

the same topics as the surveys were conducted with nine states in order to deepen our understanding of TSA impact and of the factors mediating this impact at the state level. Information for the case studies was gathered through semi-structured telephone interviews with the state assessment directors and reading specialists, and, compared to the mail surveys, the focus of the case study interviews was more narrowly on the impact of the TSAs in the area of reading. Each of the written case studies was reviewed and approved by the respondents prior to publication.

Completed mail surveys were received from 45 states, including 42 states that had participated in one or more of the TSAs and three states that had participated in none. Overall response rates were 86 percent for assessment directors, 80 percent for reading specialists, and 78 percent for mathematics specialists. In some cases, "nonresponse" resulted from the fact that states did not use reading or mathematics specialists or the positions were not currently filled. Among 11 states contacted for case studies, two declined to participate, and case studies were completed for the remaining nine.

## *Results*

### *Overall Evaluation of the TSAs*

Among assessment directors, reading specialists, and mathematics specialists in states that had participated in the TSAs, none evaluated the overall impact of the TSAs as generally negative. About half of the assessment directors and one-third of the reading and mathematics specialists evaluated the impact as generally positive, while most of the rest appraised the impact as "too limited to classify," or somewhere in between. Reading specialists were particularly likely to view the overall impact as "mixed," noting, for example, that styles of reading instruction are very much a local prerogative in many states, and that the NAEP reading materials are more aligned with practices in some districts than others.

Those who appraised the impact of the TSAs as generally positive were also likely to evaluate participation in future state NAEP assessments as very worthwhile.

### *A Climate of Change*

Changes in reading and mathematics education practices were reported by very high proportions of the states. In mathematics, virtually all of the states reported that such changes were occurring or had occurred since 1990, and the rates of reported change in reading were almost as high. The most commonly reported changes in mathematics were better alignment with NCTM standards, more emphasis on higher-order thinking skills or problem solving, development of a standards-based curriculum, and better alignment of assessment and instruction. A high proportion of states also reported the development of student performance standards in mathematics. In reading, the trends were very similar, with the highest frequency

changes including more emphasis on higher-order thinking skills, construction of meaning and reader response, better alignment with current research in reading, and the development of standards-based curricula and student performance standards.

At the same time, several of the case study states reported being hampered by funding and staff cuts and, in some instances, by a shift toward greater local control and a redefinition of the role of state departments of education. North Carolina in particular, which stood out from all other states on the basis of its creative and effective use of TSA data in programs for district and school staff, has undergone radical downsizing. At least in the short run, this will almost certainly impact its capacity to use NAEP as well in future assessment cycles.

## *The Influence of State NAEP on Instruction and Assessment*

Among the states that had both participated in the TSAs and reported education changes in progress or recently completed, about three-quarters attributed an influence to NAEP in helping to shape these changes. The perception of influence was roughly equal in both subject areas and across types of respondents. Respondents pointed particularly to the NAEP frameworks and item types as influential, while in mathematics, which reaped the benefit of being the subject of the first TSA, many more states noted the influence resulting from a general heightening of awareness caused by TSA publicity.

Reinforcing the validity of changes already contemplated or underway was the influence most frequently attributed to the TSA, but substantial percentages of states also reported using NAEP to educate local educators about planned or needed changes, as a source of ideas about what to change, and as a mechanism for helping to sell change to policy makers and legislators.

## *Problems Limiting NAEP's Utility to the States*

Not surprisingly, when asked to speak to problems limiting NAEP's utility to the states, more than half of the respondents had something to say. Very few, however, cited problems with NAEP's alignment to current research, NCTM standards, or state curricula and frameworks. Concerns with lack of alignment between NAEP and current classroom practices were also relatively rare, although they were somewhat more common in reading, where, as noted, there are more differences in opinion than in mathematics about effective instructional styles.

The biggest concerns were with the excessive lag time to reporting and the absence of any local or district results that could make the results more salient to education practitioners. Rather predictably, these latter concerns were voiced somewhat most often by the assessment directors who have had to live intimately with state NAEP during the past six years and, importantly, hold the responsibility for recruiting schools to participate. The fact that recruitment for the 1996 state NAEP was in progress during the data collection period served particularly to heighten awareness of this concern.

# Works Cited


Allen, N.L., Carlson, J.E., and Kline, D.L. *The NAEP 1994 Technical Report.* Washington, D.C.: National Center for Education Statistics, forthcoming.

American College Testing Program. *NAEP Reading Revisit: An Evaluation of the 1992 Achievement Levels Descriptions*. February 1995.

American Educational Research Association. Journal of Educational Statistics 17 (2) (Summer 1992).

Beaton, A.B. *Implementing the New Design: The NAEP 1983-84 Technical Report.* Washington, D.C.: National Center for Education Statistics, March 1987.

Benjamini, Y. and Hochberg,Y. "Controlling the false discovery rate: A practical and powerful approach to multiple testing." Journal of the Royal Statistical Society B (55) (1995).

Bohrnstedt, G.W. and Hawkins E. "Reporting the 1994 Reading Results by Achievement Levels," in *Quality and Utility: The 1994 Trial State Assessment in Reading, Background Studies*. Stanford, CA: The National Academy of Education, forthcoming.

Campbell, J.R., Donahue, P.L., Reese, C.M., and Phillips, G.W. *NAEP 1994 Reading Report Card for the Nation and the States*. Washington, D.C.: National Center for Education Statistics, January 1996.

Colvin, R.L. "Reading skills lagging in state and across U.S." *Los Angeles Times* (April 28, 1995).

Courlander, H. "Hungry Spider and the Turtle," in *The Cow-Tail Switch & Other West African Stories*. New York: Henry Holt and Company, Inc., 1947.

Cronbach, L.J. "Construct validation after thirty years," in *Intelligence: Measurement, Theory, and Practice.* Ed. R. L. Linn. Urbana, IL: University of Illinois Press, 1989.

DeStefano, L., Pearson, P.D., and Afflerbach, P. "Content Validation of the 1994 National Assessment of Educational Progress in Reading: Assessing the Relationship Between the 1994 Assessment and the Reading Framework," in *Quality and Utility: The 1994 Trial State Assessment in Reading, Background Studies.* Stanford, CA: The National Academy of Education, forthcoming.

Devito, P.J. "The Future of NAEP from the States' Perspective," in *Assessment in Transition: Monitoring the Nation's Educational Progress, Background Studies*. Stanford, CA: The National Academy of Education, forthcoming.

Donoghue, J.R. and Mazzeo, J. "Comparing IRT-based Equating Procedures for Trend Measurement in a Complex Test Design," in Educational Testing Service. Paper presented at the annual meeting of the National Council on Measurement in Education in San Francisco, CA, April 1992.

Educational Testing Service. Personal Communication with J.E. Carlson. March 15, 1996.

Finn, C.E., Jr. News Release. Washington, D.C.: National Assessment Governing Board, November 29, 1989.

Hartka, E. and McLaughlin, D.H. "A study of the Administration of the 1992 National Assessment of Educational Progress Trial State Assessment," in *The Trial State Assessment: Prospects and Realities: Background Studies.* Stanford, CA: The National Academy of Education, 1994.

Hartka, E., Perie, M., and McLaughlin, D.H. "Public School Nonparticipation Study," in *Quality and Utility: The 1994 Trial State Assessment in Reading, Background Studies.* Stanford, CA: The National Academy of Education, forthcoming.

Hartka, E. and Stancavage, F. "Perspectives on the Impact of the Trial State Assessments: State Assessment Directors, State Mathematics Specialists, and State Reading Specialists," in *Quality and Utility: The 1994 Trial State Assessment in Reading, Background Studies.* Stanford, CA: The National Academy of Education, forthcoming.

Hartka E., Yu J., and McLaughlin, D.H. "A Study of the Administration of the 1994 Trial State Assessment," in *Quality and Utility: The 1994 Trial State Assessment in Reading, Background Studies.* Stanford, CA: The National Academy of Education, forthcoming.

Hatcher, T. "S.C. students near bottom in reading." *The Post and Courier* (April 28, 1995).

Hedges, L.V. and Venesky, R.L. "The 1994 Reading Anomaly: Report to The National Academy of Education on the Drop in National Assessment of Educational Progress Main Assessment (Short-Term Trend) Scores" in *Quality and Utility: The 1994 Trial State Assessment in Reading, Background Studies.* Stanford, CA: The National Academy of Education, forthcoming.

Hessler, G.L. *Use and Interpretation of the WJ-R.* Chicago: The Riverside Publishing Company, 1993.

Jaeger, R.M. "Certification of student competence," in *Educational Measurement: Third Edition.* New York: American Council on Education and Macmillan Publishing Company, 1993.

Johnson, E.G. and Allen, N.L. *The NAEP 1990 Technical Report.* Washington, D.C.: National Center for Education Statistics, February 1992.

Johnson, E.G. and Carlson, J.E. *The NAEP 1992 Technical Report.* Washington, D.C.: National Center for Education Statistics, July 1994.

Johnson, E.G., Mazzeo, J., and Kline, D.L. *Technical Report of the NAEP 1992 Trial State Assessment Program in Mathematics.* Washington, D.C.: National Center for Education Statistics, April 1993.

Johnson, E.G., Mazzeo, J., and Kline, D.L. *Technical Report of the NAEP 1992 Trial State Assessment Program in Reading*. Washington, D.C.: National Center for Education Statistics, February 1994.

Koffler, S.L. *The Technical Report of NAEP's 1990 Trial State Assessment Program*. Washington, D.C.: National Center for Education Statistics, April 1991.

Leslie, L. and Caldwell, J. *Qualitative Reading Inventory*. La Porte, IN: Harper Collins Publishers Inc., 1990.

Linn, R.L., Koretz, D.M., Baker, E.L., and Burstein, L. *The Validity and Credibility of the Achievement Levels for the 1990 National Assessment of Educational Progress in Mathematics*. Los Angeles, CA: Center for Research on Evaluation, Standards, and Student Testing, January 1991.

Mazzeo, J., Allen, N.L., and Kline, D.L. *Technical Report of the NAEP 1994 Trial State Assessment Program in Reading*. Washington, D.C.: National Center for Education Statistics, December 1995.

McLaughlin, D.H. "Validity of the 1992 NAEP Achievement Level-Setting Process," in *Setting Performance Standards for Student Achievement: Background Studies*. Stanford, CA: The National Academy of Education, 1993.

McLaughlin, D.H. *The Problem with Item-Based Judgment Procedures for NAEP Achievement Level Setting*. January 1994. A draft document presented to Dr. John Burkett. National Center for Education Statistics, January 25, 1994.

Mitchell, J.H. "Evaluation of the 1992 Reading Framework for the National Assessment of Educational Progress," in *The Trial State Assessment: Prospects and Realities: Background Studies*. Stanford, CA: The National Academy of Education, 1993.

Morson, B. "Two-thirds of minority 4th-graders' reading skills poor." *Rocky Mountain News* (April 28, 1995).

Mullis, I.V.S., Dossey, J.A., Owen, E.H., and Phillips, G.W. *The State of Mathematics Achievement*. Washington, D.C.: National Center for Education Statistics, June 1991.

Mullis, I.V.S., Dossey, J.A., Owen, E.H., and Phillips, G.W. *NAEP 1992 Mathematics Report Card for the Nation and the States*. Washington, D.C.: National Center for Education Statistics, April 1993.

Mullis, I.V.S., Campbell, J.R., and Fartstrup, A.E. *NAEP 1992 Reading Report Card for the Nation and the States*. Washington, D.C.: National Center for Education Statistics, September 1993.

The National Academy of Education. *Assessing Student Achievement in the States*. Stanford, CA: 1992.

The National Academy of Education. *The Trial State Assessment: Prospects and Realities*. Stanford, CA: 1993.

NAEP Reading Consensus Project. *Reading Framework for the 1992 and 1994 National Assessment of Educational Progress.* Washington, D.C.: National Assessment Governing Board, 1993.

National Center on Educational Outcome. *Synthesis Report 15: Recommendations for Making Decisions about the Participation of Students with Disabilities in Statewide Assessment Programs.* Minneapolis, MN: 1994.

National Center on Educational Outcome. Personal communication with Kevin McGrew. May 23, 1995.

National Center for Education Statistics. *1990 Trial State Assessment: Summary of Participation Rates, Training, and Data Collection Activities.* Washington, D.C.: September 1990.

National Center for Education Statistics. *1992 Trial State Assessment: Summary of Participation Rates, Training, and Data Collection Activities.* Washington, D.C.: July 1992.

National Center for Education Statistics. *1994 Trial State Assessment: United States Report on Data Collection.* Washington, D.C.: October 1994.

National Center for Education Statistics. *Report on Data Collection Activities for the 1994 National Assessment of Educational Progress.* Washington, D.C.: March 1995.

National Council of Educational Measurement. Journal of Educational Measurement 29 (2) (Summer 1992).

National Education Goals Panel. *The National Education Goals Report: Building a Nation of Learners.* Washington, D.C.: 1991.

Pearson, P.D. and DeStefano, L. "Content Validation of the 1992 NAEP in Reading: Classifying Items According to the Reading Framework," in *The Trial State Assessment: Prospects and Realities: Background Studies.* Stanford, CA: The National Academy of Education, 1993.

Public Law 100-297. Part C, Section 3402: April 1988.

Public Law 103-227. Section 3, 108 Stat. 129: March 1994.

Schaver, M. "Comparing test adds confusion to KARA debate." *The Courier-Journal* (May 10, 1995).

Spencer, B.D. "School and Student Sampling in the 1994 Trial State Assessment: An Evaluation," in *Quality and Utility: The 1994 Trial State Assessment in Reading, Background Studies.* Stanford, CA: The National Academy of Education, forthcoming.

Stancavage, F., Allen, J., and Godlewski, C. "Study of Exclusion and Assessability of Students with Limited English Proficiency in the 1994 Trial State Assessment of the National Assessment of Educational Progress," in *Quality and Utility: The 1994 Trial State Assessment in Reading, Background Studies*. Stanford, CA: The National Academy of Education, forthcoming.

Stancavage, F., McLaughlin, D.H., Vergun, R., Godlewski, C., and Allen, J. "Study of Exclusion and Assessability of Students with Disabilities in the 1994 Trial State Assessment of the National Assessment of Educational Progress," in *Quality and Utility: The 1994 Trial State Assessment in Reading, Background Studies.* Stanford, CA: The National Academy of Education, forthcoming.

Stufflebeam, D.M., Jaeger, R.M., and Scriven, M. *Summative Evaluation of the National Assessment Governing Board's Inaugural 1990-91 Effort to Set Achievement Levels on the National Assessment of Educational Progress.* Washington, D. C.: National Assessment Governing Board, 1991.

Thorndike, R.L. "Item and Score Conversion by Pooled Judgment," in *Test Equating.* Ed. P. W. Holland and D. B. Rubin. New York: Academic Press, 1982.

Turgeon, A. "Rhode Island's fourth-graders better reading scores." *Providence Journal-Bulletin* (April 29, 1995).

U.S. General Accounting Office. *National Assessment Technical Quality.* GAO/PEMD-92-22R. Washington, D.C.: March 1992.

U. S. General Accounting Office. *Educational Achievement Standards: NAGB's Approach Yields Misleading Interpretations.* GAO/PEMD-93-12. Washington, D.C.: June 1993.

White, B. "Scores mediocre everywhere; Georgia's students rank 39th." *Atlanta Constitution* (April 28, 1995).

Williams, P.L., Reese, C.M., Campbell, J.R., Mazzeo, J., and Phillips, G.W. *NAEP 1994 Reading: A First Look.* Washington, D.C.: National Center for Education Statistics, April 1995.

Williams, V.S.L., Jones, L.V., and Tukey, J.W. *Controlling error in multiple comparisons with special attention to the National Assessment of Educational Progress: Technical Report No. 33.* Research Triangle Park, NC: National Institute of Statistical Sciences, 1994.

Woodcock, R.W. and Johnson, M.B. *WJ-R Tests of Achievement: Standard and Supplemental Batteries.* Chicago: The Riverside Publishing Company, 1989.

Young, S. "Scores belie proficiency, says survey: 1/3 of Maine's 4th graders considered skilled readers." *Bangor Daily News* (April 28, 1995).

# List of Abbreviations

| | |
|---|---|
| AP | Advanced Placement |
| ACT | American College Testing |
| AIR | American Institutes for Research |
| CCD | Common Core of Data |
| CCSSO | Council of Chief State School Officers |
| DODEA | Department of Defense Education Activity |
| ETS | Educational Testing Service |
| GAO | General Accounting Office |
| IEP | Individualized Education Plan |
| IRT | Item Response Theory |
| LEP | Limited English Proficiency |
| NAE | The National Academy of Education |
| NAGB | National Assessment Governing Board |
| NCEO | National Center on Educational Outcome |
| NCES | National Center for Education Statistics |
| NCS | National Computer Systems |
| NCEST | National Council of Education Standards & Testing |
| NCTM | National Council of Teachers of Mathematics |
| NEGP | National Education Goals Panel |
| NELS | National Education Longitudinal Study |
| OBEMLA | Office of Bilingual Education and Minority Language Affairs |
| OERI | Office of Educational Research and Improvement |
| QED | Quality Education Data |
| SAT | Scholastic Assessment Test |
| SEA | State Education Agency |
| TSA | Trial State Assessment |
| WJBRC | Woodcock Johnson Broad Reading Cluster |

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement (OERI)
Educational Resources Information Center (ERIC)

# NOTICE

## REPRODUCTION BASIS

☐ This document is covered by a signed "Reproduction Release (Blanket)" form (on file within the ERIC system), encompassing all or classes of documents from its source organization and, therefore, does not require a "Specific Document" Release form.

☑ This document is Federally-funded, or carries its own permission to reproduce, or is otherwise in the public domain and, therefore, may be reproduced by ERIC without a signed Reproduction Release form (either "Specific Document" or "Blanket").